FIT5197, 2020 Semester 2, Formula Sheet

Daniel Schmidt, Wray Buntine, Levin Kuhlmann

November 7, 2020

1 Sample Statistics

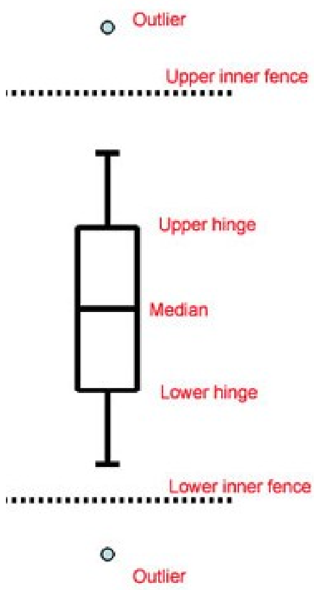
- quartiles, percentiles, etc.: given n data points, rank them in increasing value to get $x_1, ..., x_n$
  - median, if n is odd, given by $x_{(n+1)/2}$, if n is even given by $\frac{1}{2}(x_{n/2} + x_{n/2+1})$
  - quartiles, Q_1 or Q_3 is given by $Q_k = x_p + \frac{q}{4}(x_{p+1} - x_p)$ where $p = floor((k(n+1))/4)$ and $q = (k(n+1)) \bmod 4$
  - percentiles, $P_k = x_p + \frac{q}{100}(x_{p+1} - x_p)$ where $p = floor((k(n+1))/100)$ and $q = (k(n+1)) \bmod 100$
- measures of spread for n data points, $\boldsymbol{x} = (x_1, ..., x_n)$
  - Sample variance, $\text{var}(\boldsymbol{x}) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
  - Sample standard deviation, s_x
  - Range $= \max_{i=1}^n x_i - \min_{i=1}^n x_i$
  - (inter-quartile range) IQR $= Q_3 - Q_1$
- Sample covariance

$$q_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Sample correlation coefficient

$$r_{xy} = \frac{q_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- boxplots
  - represents numerical data through quartiles
  - the lower hinge is Q_1, upper hinge is Q_3
  - a lower whisker is drawn at minimum data value greater than the lower inner fence is $Q1 - 1.5 \times \text{IQR}$ (which itself is usually not draw)
  - upper whisker is drawn at maximum data value less than the upper inner fence is $Q3 + 1.5 \times \text{IQR}$ (which itself is usually not draw)
  - outliers are highlighted outside these two fences



2 Probability

- probability axioms of Kolmogorov:
  1. for any event A, $0 \leq p(A) \leq 1$
  2. $p(\Omega) = 1$, where Ω is the universal set, the set of everything
  3. for mutually exclusive events A_1, ... A_n $p(A_1 \cup A_2 ... \cup A_n) = \sum_{i=1}^{n} p(A_i)$
- other probability identities for the domain $\mathcal{X} \times \mathcal{Y}$ where A, B are any events:
  - complement rule, $p(\overline{A}) = 1 - p(A)$
  - product rule, $p(B \cap A) = p(B|A)p(A)$
  - sum rule, $p(A) = \sum_{x \in \mathcal{X}} p(x \cap A)$
  - conditional, $p(B|A) = \frac{p(B \cap A)}{p(A)}$ when $p(A) > 0$
  - Bayes theorem, $p(x|A) = \frac{p(A|x)p(x)}{\sum_{x \in \mathcal{X}} p(A|x)p(x)}$

- for continuous random variables a **probability density function** (PDF) $p(x)$ on domain $\mathcal{X}$ satisfies

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X}$$

  and

$$\int_{\mathcal{X}} p(x)dx = 1$$

- for continuous random variables, the probability $X \in A$, where $A \subset \mathcal{X}$ is

$$p(X \in A) = \int_A p(x)dx.$$

- for $\mathcal{X}$ a single dimension, then define the **cumulative density function** (CDF), $P(x)$, in terms of the the PDF $p(x)$ as

$$P(x) = \int_{y<x} p(y)\mathrm{d}\, y$$

  and the **quantile function** $Q(x)$ as

$$Q(x) = P^{-1}(x)$$

  this is well defined when $p(x) > 0$.

- Let the random variable pair (X, Y) be from domain $\mathcal{X} \times \mathcal{Y}$. We say X and Y are **independent** if any of the following three (equivalent) conditions hold for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$\begin{array}{rrcll} (I) & p(X{=}x|Y{=}y) & = & p(X{=}x) & \text{when } p(Y{=}y) > 0 \\ (II) & p(Y{=}y|X{=}x) & = & p(Y{=}y) & \text{when } p(X{=}x) > 0 \\ (III) & p(Y{=}y \cap X{=}x) & = & p(X{=}x)p(Y{=}y) & \end{array}$$

3 Expected Values

- if X has domain $\mathcal{X}$, **expectation** and **variance** of $f(X)$:

$$\begin{aligned} \mathbb{E}\left[f(X)\right] &= \sum_{x \in \mathcal{X}} p(x) f(x) \\ \mathbb{V}\left[f(X)\right] &= \mathbb{E}\left[(f(X) - \mathbb{E}\left[f(X)\right])^2\right] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}\left[f(X)\right]^2 \end{aligned}$$

  with integral replacing sum for continuous RVs

- some useful rules for RVs X, Y and constant c
  - $\mathbb{E}[f(X)+g(Y)] = \mathbb{E}[f(X)] + \mathbb{E}[g(Y)]$
  - $\mathbb{E}[cf(X)] = c\mathbb{E}[f(X)]$
  - $\mathbb{V}[cf(X)] = c^2\mathbb{V}[f(X)]$
- if X, Y are independent RVs
  - $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$
  - $\mathbb{V}[f(X)+g(Y)] = \mathbb{V}[f(X)] + \mathbb{V}[g(Y)]$
- **Chebyshev's inequality**: if X is a RV with mean μ and variance σ^2, then for any $k > 0$
$$p\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$
- **Chebyshev's inequality for samples**: for a sample $S = \{x_1, ..., x_N\}$ of variable X with mean $\overline{x}$ and sample standard deviation s_x, then for any $k > 0$
$$\left|\left\{x_i \,:\, \frac{|x_i - \overline{x}|}{s_x} \geq k\right\}\right| \leq \frac{N}{k^2}$$
that is, the number of data points at least $k\,s_x$ from the mean is no more than $\frac{N}{k^2}$.
- **Weak law of large numbers**: let $X_1, \ldots, X_n$ be RVs with $\mathbb{E}[X_i] = \mu$; then for any $\varepsilon > 0$
$$p\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

4 Distributions

- for the **Gaussian or normal distribution**, denoted $N(\mu, \sigma^2)$
$$p(x \,|\, \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
and has the properties

- $\mathbb{E}[x] = \mu$ and $\mathbb{V}[x] = \sigma^2$
- the mode and the median are the same as the mean
- if the curve for $p(x \mid 0, 1)$ is shifted to the right by μ and scaled by $1/\sigma$, one gets the curve for $p(x \mid \mu, \sigma^2)$

- the **discrete uniform distribution** models discrete RVs denoted $U(a, b)$ and follows
$$\mathbb{P}(X = k \mid a, b) = \frac{1}{b - a + 1}$$
where $X \in \{a, \ldots, b\}$ with $b \geq a$, and has properties
  - $\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathbb{V}[X] = \frac{(b-a+1)^2-1}{12}$

- the **continuous uniform distribution** models continuous RVs denoted $U(a, b)$ with pdf
$$p(x \mid a, b) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x > b \end{cases}$$
where $a > b$ and
  - The quantity a determines the start of the distribution
  - The quantity $w = b - a$ is the width of the distribution
  - $\mathbb{E}[X] = \frac{a+b}{2} = a + \frac{w}{2}$ and $\mathbb{V}[X] = \frac{(b-a)^2}{12} = \frac{w^2}{12}$

- the **Bernoulli distribution** models discrete, binary RVs, i.e., $\mathcal{X} = \{0, 1\}$, denoted $\mathrm{B}e(\theta)$,
$$p(X = 1 \mid \theta) = \theta, \ \theta \in [0, 1]$$
so that the parametric probability distribution is
$$p(x \mid \theta) = \theta^x (1 - \theta)^{(1-x)}$$
and has properties
  - $\mathbb{E}[x] = \theta$ and $\mathbb{V}[x] = \theta(1 - \theta)$

- the **binomial distribution** describes the probability of getting x successful outcomes in n Bernoulli trials with probability of success θ, denoted $\text{bin}(\theta, n)$, and $x \in \{0, 1, ..., n\}$,

$$p(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

  and has properties

  - $\mathbb{E}[x] = n\theta$ and $\mathbb{V}[x] = n\theta(1-\theta)$

- the **Poisson distribution** with rate parameter λ is the number of events x occurring, for $\mathcal{X} = \{0\} \cup \mathcal{N}$, denoted $\text{Pois}(\lambda)$,

$$p(x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

  and has properties

  - $\mathbb{E}[x] = \lambda$ and $\mathbb{V}[x] = \lambda$
  - if $X \sim \text{Pois}(\lambda_X)$ and $Y \sim \text{Pois}(\lambda_Y)$ then $(X+Y) \sim \text{Pois}(\lambda_X + \lambda_Y)$
  - $\text{bin}(\theta, n) \approx \text{Pois}(n\theta)$ for $n \gg 1$ and $n\theta$ small

- Note the Central Limit Theorem (CLT) has been moved to section 6 of this document.

5 Estimation

- have a sample $\boldsymbol{x}$; let $\hat{\theta}(\boldsymbol{x})$ be a point estimate for model parameter θ; then $\hat{\theta}(\boldsymbol{x})$ is **unbiased** if $\mathbb{E}_{\boldsymbol{x}}\left[\hat{\theta}(\boldsymbol{x})\right] = \theta$, where the expectation is taken over samples $\boldsymbol{x}$

- the **bias** of the estimator is

$$b_\theta(\hat{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\hat{\theta}(\boldsymbol{x})\right] - \theta$$

- the **variance** of the estimator is

$$\mathbb{V}_\theta\left[\hat{\theta}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{\theta}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}\left[\hat{\theta}(\boldsymbol{x})\right]\right)^2\right]$$

- the **mean square error** (MSE) of the estimator is

$$\mathbb{MSE}_\theta\left[\hat{\theta}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\left(\hat{\theta}(\boldsymbol{x}) - \theta\right)^2\right] = b_\theta(\hat{\theta})^2 + \mathbb{V}_\theta\left[\hat{\theta}\right]$$

- for sample $\boldsymbol{x}$ of size n distributed as $N(\mu, \sigma)$ the **sum of squared errors** (SSE) of mean estimate μ is given by

$$\mathrm{SSE}(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2$$

  and the point estimate $\hat{\mu}$ minimising SSE is the mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- the method of maximum likelihood says we should use the model that assigns the greatest probability to the data we have observed; formally, the **maximum likelihood estimator** (MLE) is found by solving

$$\hat{\Theta} = \arg\max_{\Theta}\{p(\boldsymbol{x} \mid \Theta)\}$$

  where $p(\boldsymbol{x} \mid \Theta)$ is called the **likelihood function**

- use $L(\boldsymbol{x} \mid \Theta)$ to denote the **negative log-likelihood**, $\log 1/p(\boldsymbol{x} \mid \Theta)$
- for sample $\boldsymbol{x}$ of size n distributed as $N(\mu, \sigma)$

$$L(\boldsymbol{x} \mid \mu, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathrm{SSE}(\mu)$$

  from this we get

  - $\hat{\mu}_{ML}$ is the mean, same as when using the SSE
  - the MLE for the variance is

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

    this is however biased, an unbiased estimate is

$$\hat{\sigma}^2_u = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

- the MLE estimates for λ of the Poisson and θ of the Bernoulli is also the mean
- the MLE estimates for θ of the binomial, $\text{bin}(\theta, m)$, using sample $\boldsymbol{x}$ of size n is
$$\hat{\theta}_{ML} = \frac{1}{nm}\sum_{i=1}^{n} x_i$$
- let $\boldsymbol{x}$ be a sample of size n from a Gaussian population with mean μ and variance σ^2, and let m be the mean and s^2 be the sample variance:
  - m is Gaussian with mean, variance $\left(\mu, \frac{\sigma^2}{n}\right)$
  - $\sqrt{n}(m-\mu)/s$ is Student's t with $n-1$ degrees of freedom
  - these can be used to develop confidence bounds or hypothesis tests for μ and σ^2 respectively
- the **Student's t distribution** with n degrees of freedom, denoted Student–t(n), has the following properties:
  - it looks like a standard normal as $n \to \infty$
  - is symmetric about 0
  - has mean $\mathbb{E}_{\text{Stu}_n}[X] = 0$ for $n > 1$
    - mean undefined for $n = 1$
  - has variance $\mathbb{V}_{\text{Stu}_n}[X] = \frac{n}{n-2}$ for $n > 2$
    - variance undefined for $n \leq 2$

6 CLT and Confidence Intervals

- **Central Limit Theorem** (CLT): have distribution with mean μ and variance σ^2, and sample n identical RVs $X_1, ..., X_n$ from it; then the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately distributed as $N\left(\mu, \frac{1}{n}\sigma^2\right)$ for large n. Likewise the sample sum $\sum_{i=1}^{n} X_i$ is approximately distributed as $N\left(n\mu, n\sigma^2\right)$ for large n.
- examples of the CLT
  - it is exact in the case of the Gaussian

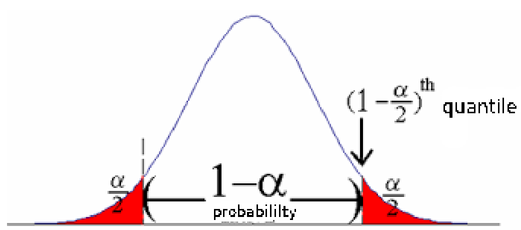
- for the binomial, $\text{bin}(\theta, n) \approx N(n\theta, n\theta(1-\theta))$ for $n \gg 1$ and θ not near 0 or 1
- for the Poisson, $\text{Pois}(\lambda) \approx N(\lambda, \lambda)$ for $\lambda \gg 1$

- let X have the CDF $P(X)$, and $Q(p) = P^{-1}(p)$ is the corresponding quantile function; then the $(1-\alpha)$ **two-sided confidence interval** for X is given by

$$[Q(\alpha/2),\, Q(1-\alpha/2)]$$

- consider the case for $Z \sim N(0,1)$:
  - let $Z_{1-\alpha/2}$ denotes the upper $\alpha/2$ quantile for $N(0,1)$
  - we are $1-\alpha$ confident $Z \sim N(0,1)$ falls inside $(-Z_{1-\alpha/2}, Z_{1-\alpha/2})$
  - $[-Z_{1-\alpha/2}, Z_{1-\alpha/2}]$ is called a (two-sided) confidence interval for $N(0,1)$

  this is depicted in the unshaded part of the curve:



- let X have the CDF $P(X)$, and $Q(p) = P^{-1}(p)$ is the corresponding quantile function; then the $(1-\alpha)$ **one-sided lower confidence interval** for X is given by

$$[-\infty,\, Q(1-\alpha)]$$

  and the $(1-\alpha)$ **one-sided upper confidence interval** for X is given by

$$[Q(\alpha),\, \infty]$$

- assume dataset of count n with mean $\bar{X}$ and sample variance S^2:

assumptions	parameter	interval
Gaussian, σ^2 known	μ	$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Gaussian, σ^2 unknown	μ	$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$

- assume dataset of count n with mean $\bar{X}$ and sample variance S^2 and a second dataset of count m with mean $\bar{Y}$ and sample variance T^2:

assumptions	parameter	interval
Gaussian, σ_1^2, σ_2^2 known	$\mu_1 - \mu_2$	$\bar{X} - \bar{Y} \pm Z_{\alpha/2} \sqrt{\sigma_1^2/n + \sigma_2^2/m}$
Gaussian, $\sigma_1^2 = \sigma_2^2$ unknown but equal	$\mu_1 - \mu_2$	$\bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\, S_P \sqrt{\frac{1}{n} + \frac{1}{m}}$ for $S_P^2 = \frac{(n-1)S^2+(m-1)T^2}{n+m-2}$
Gaussian, $\sigma_1^2 \neq \sigma_2^2$ unknown, using CLT	$\mu_1 - \mu_2$	use 1st case for $\sigma_1^2 = S^2$, $\sigma_2^2 = T^2$, assuming n, m are large

- for Poisson, assume dataset of count n with mean $\hat{X}$; for Bernoulli, assume dataset of count n with mean $\hat{p}$; also a 2nd dataset of count m with mean $\hat{q}$;

assumptions	parameter	interval
Poisson, λ unknown, using CLT	λ	$\hat{X} \pm Z_{\alpha/2} \sqrt{\hat{X}/n}$
Bernoulli, θ unknown, using CLT	θ	$\hat{p} \pm Z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$
Bernoulli, θ_1, θ_2 unknown, using CLT	$\theta_1 - \theta_2$	$\hat{p} - \hat{q} \pm Z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n + \hat{q}(1-\hat{q})/m}$

7 Hypothesis Tests

- given an arbitrary **test statistic** x with CDF $P(X)$ (i.e. x could be z or t), then the p-value is given by

$$p = \begin{cases} 2P(-|x|) & \text{if null hypothesis is equality} \\ 1 - P(x) & \text{if null hypothesis involves } \leq \\ P(x) & \text{if null hypothesis involves } \geq \end{cases}$$

- assume dataset of count n with mean $\bar{X}$ and sample variance S^2:

assumptions	null-hypo.	test statistic
Gaussian, σ^2 known	μ_0	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
Gaussian, σ^2 unknown	μ_0	$t_{n-1} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$

- assume dataset of count n with mean $\bar{X}$ and sample variance S^2 and a second dataset of count m with mean $\bar{Y}$ and sample variance T^2:

assumptions	null-hypo.	test statistic
Gaussian, σ_1^2, σ_2^2 known	$\Delta\mu_0$	$Z = \frac{\bar{X}-\bar{Y}-\Delta\mu_0}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}$
Gaussian, $\sigma_1^2 = \sigma_2^2$ unknown but equal	$\Delta\mu_0$	$t_{n+m-2} = \frac{\bar{X}-\bar{Y}-\Delta\mu_0}{S_P\sqrt{\frac{1}{n}+\frac{1}{m}}}$ for $S_P^2 = \frac{(n-1)S^2+(m-1)T^2}{n+m-2}$
Gaussian, $\sigma_1^2 \neq \sigma_2^2$ unknown, using CLT	$\Delta\mu_0$	use 1st case for $\sigma_1^2 = S^2$, $\sigma_2^2 = T^2$, assuming n, m are large

- for Poisson, assume dataset of count n with mean $\hat{X}$; for Bernoulli, assume dataset of count n with mean $\hat{p}$; also a 2nd dataset of count m with mean $\hat{q}$; all using the CLT so require large samples (n, m)

assumptions	null-hypo.	test statistic
Poisson, λ unknown, using CLT	λ_0	$Z = \frac{\hat{X}-\lambda_0}{\sqrt{\lambda_0/n}}$
Bernoulli, θ unknown, using CLT	θ_0	$Z = \frac{\hat{p}-\theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$
Bernoulli, θ_1, θ_2 unknown, using CLT	$\Delta\theta_0$	$Z = \frac{\hat{p}-\hat{q}-\Delta\theta_0}{\sqrt{\hat{p}(1-\hat{p})/n+\hat{q}(1-\hat{q})/m}}$ If $\Delta\theta_0 = 0$ this reduces to $Z = \frac{\hat{p}-\hat{q}}{\sqrt{\hat{r}(1-\hat{r})(1/n+1/m)}}$ where $\hat{r} = \frac{n\hat{p}+m\hat{q}}{n+m}$

8 Linear Regression

- simple least squares model has $\mathbb{E}\left[y_i \mid x_i\right] = \beta_0 + \beta_1 x_i$ and has a residual sum of squares

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- various intermediate formula are used to calculate quantities including the sums

$$\begin{aligned}
\mathrm{SS}_{XX} &= \sum_{i=1}^{n}(x_i - \overline{X})^2 = n\left(\overline{X^2} - \overline{X}^2\right) \\
\mathrm{SS}_{XY} &= \sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = n\left(\overline{XY} - \overline{X}\,\overline{Y}\right) \\
\mathrm{SS}_{YY} &= \sum_{i=1}^{n}(y_i - \overline{Y})^2 = n\left(\overline{Y^2} - \overline{Y}^2\right)
\end{aligned}$$

- with this the RSS can be minimised using the solution for β_1 of

$$\hat{\beta}_1 = \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_{XX}} = \frac{\overline{XY} - \overline{X}\;\overline{Y}}{\overline{X^2} - \overline{X}^2}$$

and the solution for β_0 of

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = \frac{\overline{Y}\,\overline{X^2} - \overline{XY}\,\overline{X}}{\overline{X^2} - \overline{X}^2}$$

- giving an RSS at the minimum of

$$\text{RSS}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\text{SS}_{YY}\text{SS}_{XX} - \text{SS}_{XY}^2}{\text{SS}_{XX}} = \text{SS}_{YY} - \text{SS}_{XX}\hat{\beta}_1^{\,2}$$

- if we use the probability model $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ then the log-likelihood becomes

$$L(\boldsymbol{x}, \boldsymbol{y} \,|\, \beta_0, \beta_1, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{\text{RSS}(\beta_0, \beta_1)}{2\sigma^2}$$

- minimising this gives the same solution to β_0, β_1 as before and an estimator for σ^2

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\text{RSS}(\hat{\beta}_0, \hat{\beta}_1)$$

  plus an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2_u = \frac{1}{n-2}\text{RSS}(\hat{\beta}_0, \hat{\beta}_1)$$

- moreover, the following statistics can be used to develop confidence intervals or hypothesis tests

$$\begin{aligned}
\frac{1}{\sigma^2}\text{RSS}(\hat{\beta}_0, \hat{\beta}_1) &\sim \chi^2_{n-2} \\
\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{\overline{X^2}}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_0 - \beta_0) &\sim \text{Student–t}(n-2) \\
\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_1 - \beta_1) &\sim \text{Student–t}(n-2)
\end{aligned}$$

- a measure of quality for the linear regression is the $\mathbf{R^2}$ value computed as

$$R^2 = 1 - \frac{RSS}{\text{SS}_{YY}} = \frac{\text{SS}_{XY}^2}{\text{SS}_{XX}\text{SS}_{YY}} = r^2_{XY}$$

  which is in $[0, 1]$, 1 for a perfect zero error fit, 0 for pure noise, and higher for better quality fit

- for **multi-linear regression**, the prediction instead becomes

$$\mathbb{E}\left[y_i \,|\, x_{i,1}, \ldots, x_{i,p}\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

and residual sum of squares (RSS) of

$$\text{RSS}(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

- the **design matrix** given by $\mathbf{X}$ of predictors:

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_p) = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix},$$

has corresponding parameters $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ yielding the prediction

$$\mathbb{E}\left[y_i \,|\, \boldsymbol{x}_i\right] = \mathbf{X}\boldsymbol{\beta}$$

- minimising $\text{RSS}(\boldsymbol{\beta})$ has solution

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}) \\ \text{RSS}(\hat{\boldsymbol{\beta}}) &= \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^T(\mathbf{X}^T\mathbf{Y}) \end{aligned}$$

- if we use the probability model $y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)$ then the log-likelihood becomes

$$L(\mathbf{X}, \boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{\text{RSS}(\boldsymbol{\beta})}{2\sigma^2}$$

- minimising this gives the same solution to $\boldsymbol{\beta}$ as before and an estimator for σ^2

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\text{RSS}(\hat{\boldsymbol{\beta}})$$

plus an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2_u = \frac{1}{n-p-1}\text{RSS}(\hat{\boldsymbol{\beta}})$$

9 Classification and Clustering

- probability prediction formula for **naïve Bayes classifier** is

$$P(y \mid x_1, \ldots, x_p) = \frac{P(y) \prod_{j=1}^{p} P(x_j \mid y)}{P(x_1, \ldots, x_p)}$$

  where the denominator is a constant so can be found by normalising the renumerator.

- point estimation is done by estimating the probabilities $P(Y = y)$ and $P(X_j = x_j \mid Y = y)$ for all entries of the tables

- probability prediction formula for **logistic regression classifier** is expressed using the **logistic function**

$$p(Y_i = 1 \mid x_{i,1}, ..., x_{i,p}) = \frac{1}{1 + \exp(-\eta_i)}$$

  where

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

  so that the **log-odds** given by

$$\log \frac{p(Y_i = 1 \mid x_{i,1}, ..., x_{i,p})}{p(Y_i = 0 \mid x_{i,1}, ..., x_{i,p})} = \eta_i$$

- the parameters $(\beta_0, \beta_1, ...\beta_p)$ are fit using optimising routines on the log likelihood

10 More Classification

- $log_2(x) = log_c(x)/log_c(2)$ where c is any constant

- define **entropy** (to base 2 by default)

$$H(X) = \mathbb{E}\left[\log_2 1/p(X)\right]$$

- define **condtional entropy** (to base 2 by default)

$$H(X|Y) = \sum_y p(Y{=}y)H(X|Y{=}y)$$

where $H(X|Y{=}y)$ is the entropy of the conditional distribution $p(X|Y{=}y)$.

- some properties of entropy where X has discrete domain $\mathcal{X}$:
  - if $\mathcal{X}$ of finite dimension K, then $0 \leq H(X) \leq K$
  - if $H(X) = 0$ then $p(X{=}x) = 1$ for some $x \in \mathcal{X}$
- some useful rules for RVs X, Y
  - $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- if X, Y are independent RVs
  - $H(X, Y) = H(X) + H(Y)$
- Information gain for predictor RV X and target RV Y is defined as
  - $I.G.(Y, X) = H(Y) - H(Y|X)$

11 Simulation

- To obtain an inverse $x = f^{-1}(y)$ of the function $y = f(x)$ you need to make sure that for every value of x there is only one value of $y = f(x)$ for the domain of x being considered and then rearrange $y = f(x)$ so that x can be expressed as a function of y. Then also the domain and range of $y = f(x)$ becomes the range and domain of $x = f^{-1}(y)$, respectively.
- Make sure you memorise the inverse sampling and rejection sampling algorithms or you will get to the exam and read this and say ”oh, bugger”.

12 Tables for Standard Normal

Tables from http://www.z-table.com/ on the next 2 pages. One table for z-values less than 0 and one table for z-values greater than 0 to help you find $p = F(z)$.

Table for z-values less than 0.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table for z-values greater than 0.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

13 Table for Student t

Table from http://www.ttable.org/. Provides critical t-values for specific significance values for one- and two-sided t-tests.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

14 Calculus

CALCULUS

DEFINITE INTEGRAL DEFINITION

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x$$

where $\Delta x = \frac{b-a}{n}$ and $x_k = a + k\Delta x$

FUNDAMENTAL THEOREM OF CALCULUS

$$\int_a^b f(x)dx = [F(x)]_a^b = F(b) - F(a)$$

where f is continuous on $[a,b]$ and $F' = f$

INTEGRATION PROPERTIES

$$\int_a^b cf(x)dx = c\int_a^b f(x)dx$$

$$\int_a^b f(x) \pm g(x)dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx$$

$$\int_a^a f(x)dx = 0 \text{ and } \int_a^b f(x)dx = -\int_b^a f(x)dx$$

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$$

COMMON INTEGRALS

$$\int k\ dx = kx + C$$

$$\int x^n\ dx = \frac{1}{n+1}x^{n+1} + C, n \neq -1$$

$$\int x^{-1}\ dx = \int \frac{1}{x}\ dx = \ln|x| + C$$

$$\int \frac{1}{ax+b}\ dx = \frac{1}{a}\ln|ax+b| + C$$

$$\int \ln(x)\ dx = x\ln(x) - x + C$$

$$\int e^x\ dx = e^x + C$$

$$\int \cos x\ dx = \sin x + C$$

$$\int \sin x\ dx = -\cos x + C$$

$$\int \sec^2 x\ dx = \tan x + C$$

$$\int \sec x \tan x\ dx = \sec x + C$$

$$\int \csc x \cot x\ dx = -\csc x + C$$

$$\int \csc^2 x\ dx = -\cot x + C$$

$$\int \tan x\ dx = \ln|\sec x| + C$$

$$\int \sec x\ dx = \ln|\sec x + \tan x| + C$$

$$\int \frac{1}{a^2+u^2}\ dx = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right) + C$$

$$\int \frac{1}{\sqrt{a^2-u^2}}\ dx = \sin^{-1}\left(\frac{u}{a}\right) + C$$

Basic Differentiation Rules for Elementary Functions

1. $\frac{d}{dx}[cu] = cu'$
2. $\frac{d}{dx}[u \pm v] = u' \pm v'$
3. $\frac{d}{dx}[uv] = uv' + vu'$
4. $\frac{d}{dx}\left[\frac{u}{v}\right] = \frac{vu' - uv'}{v^2}$
5. $\frac{d}{dx}[c] = 0$
6. $\frac{d}{dx}[u^n] = nu^{n-1}u'$
7. $\frac{d}{dx}[x] = 1$
8. $\frac{d}{dx}[|u|] = \frac{u}{|u|}(u'), \quad u \neq 0$
9. $\frac{d}{dx}[\ln u] = \frac{u'}{u}$
10. $\frac{d}{dx}[e^u] = e^u u'$
11. $\frac{d}{dx}[\log_a u] = \frac{u'}{(\ln a)u}$
12. $\frac{d}{dx}[a^u] = (\ln a)a^u u'$
13. $\frac{d}{dx}[\sin u] = (\cos u)u'$
14. $\frac{d}{dx}[\cos u] = -(\sin u)u'$
15. $\frac{d}{dx}[\tan u] = (\sec^2 u)u'$
16. $\frac{d}{dx}[\cot u] = -(\csc^2 u)u'$
17. $\frac{d}{dx}[\sec u] = (\sec u \tan u)u'$
18. $\frac{d}{dx}[\csc u] = -(\csc u \cot u)u'$
19. $\frac{d}{dx}[\arcsin u] = \frac{u'}{\sqrt{1-u^2}}$
20. $\frac{d}{dx}[\arccos u] = \frac{-u'}{\sqrt{1-u^2}}$
21. $\frac{d}{dx}[\arctan u] = \frac{u'}{1+u^2}$
22. $\frac{d}{dx}[\text{arccot}\ u] = \frac{-u'}{1+u^2}$
23. $\frac{d}{dx}[\text{arcsec}\ u] = \frac{u'}{|u|\sqrt{u^2-1}}$
24. $\frac{d}{dx}[\text{arccsc}\ u] = \frac{-u'}{|u|\sqrt{u^2-1}}$

INTEGRATION BY SUBSTITUTION

$$\int_a^b f(g(x))\, g'(x)dx = \int_{g(a)}^{g(b)} f(u)du$$

where $u = g(x)$ and $du = g'(x)dx$

Basic Derivatives Rules

Constant Rule: $\frac{d}{dx}(c) = 0$

Constant Multiple Rule: $\frac{d}{dx}[cf(x)] = cf'(x)$

Power Rule: $\frac{d}{dx}(x^n) = nx^{n-1}$

Sum Rule: $\frac{d}{dx}[f(x)+g(x)] = f'(x)+g'(x)$

Difference Rule: $\frac{d}{dx}[f(x)-g(x)] = f'(x)-g'(x)$

Product Rule: $\frac{d}{dx}[f(x)g(x)] = f(x)g'(x)+g(x)f'(x)$

Quotient Rule: $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}$

Chain Rule: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$