5 Probability and Statistics

5.1 Probability

- Two events A and B are **mutually exclusive** if $P(A \cap B) = 0$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- A and B are **independent** if $P(A|B) = P(A)$, so if they are independent $P(A \cap B) = P(A)P(B)$.

5.2 Discrete random variables

- $P(X = x)$ is the probability that the r.v X will assume a value of x.
- A discrete r.v can assume a countable number of values.
- For a d.r.v taking values $x_1, x_2, x_3, ..., x_n$, the **probability distribution** is defined as $P(X = x_i)$, such that:

$$0 \leq P(X = x_i) \leq 1 \qquad \text{and} \qquad \sum_{\text{all } i} P(X = x_i) = 1$$

- The expectation of a d.r.v:
  $E(X) = \mu = \sum xP(X = x)$
  $E(g(X)) = \sum g(x)P(X = x)$
  $E(a) = a$
  $E(aX \pm b) = aE(X) \pm b$
  $E(X \pm Y) = E(X) \pm E(Y)$
- The variance of a d.r.v:
  $\text{Var}(X) = \sigma^2 = E((x - \mu)^2) = E(X^2) - [E(X)]^2$
  $\text{Var}(a) = 0$
  $\text{Var}(aX + b) = a^2\text{Var}(X)$
  $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ (only if X and Y are independent)
- Note: never subtract variance.

5.3 Discrete distributions

<u>The Binomial distribution</u>

$$X \sim B(n, p) \qquad P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad E(X) = np \qquad \text{Var}(X) = npq$$

- There are n independent trials, two possible outcomes (either 'success' or 'failure'), with constant probability of success p, X is the number of 'successes'.
- The Binomial distribution is a combination of n Bernoulli trials.
- For $P(X \leq x)$, we find $P(X = 0) + P(X = 1) + P(X = 2) + ... + P(X = x)$.

<u>The Poisson distribution</u>

$$X \sim Po(\lambda) \qquad P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad E(X) = \text{Var}(X) = \lambda$$

- For a random variable in time or space, if there is no chance of simultaneous events, the events are independent, and the events have a constant probability of occuring, it is a Poisson process.
- λ is the parameter, and defines the number of events in a given time/space.
- If $X \sim Po(\lambda)$ and $Y \sim Po(\mu)$, then $X + Y \sim Po(\lambda + \mu)$.

The Geometric distribution

$$X \sim Geo(p) \qquad P(X=x) = pq^{x-1},\ x \geq 1 \qquad E(X) = \frac{1}{p} \qquad \text{Var}(X) = \frac{q}{p^2}$$

If we perform a series of independent trials with a probability p of success, X is the number of trials up to and including the first success.

$$\begin{aligned} P(X > x) &= P(X = x+1) + P(X = x+2) + \dots \\ &= pq^x + pq^{x+1} + pq^{x+2} + \dots \\ &= pq^x(1 + q + q^2 + ...) = pq^x(\frac{1}{1-q}) = q^x \end{aligned}$$

$$P(X > a+b | X > a) = P(X > b) = q^b$$

The Negative Binomial distribution

$$X \sim NB(r,p) \qquad P(X=x) = \binom{x-1}{r-1} p^r q^{x-r},\ r \geq 1,\ x \geq 1 \qquad E(X) = \frac{r}{p} \qquad \text{Var}(X) = \frac{rq}{p^2}$$

- X is the number of trials needed to achieve r successes.
- The Negative Binomial distribution is just a combination of r geometric trials.

5.4 Continuous random variables and CDFs

- Instead of probability distributions, we have probability density functions (PDFs), denoted by $f(x)$.
  - $f(x) \geq 0$ for all $x \in \mathbb{R}$
  - $\int_{-\infty}^{\infty} f(x)\ dx = 1$
- Continuous $\implies$ uncountable, so $P(X = x) = 0$. Therefore, $\geq$ or $>$ is irrelevant.

$$P(a < X < b) = \int_a^b f(x)\ dx$$

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)\ dx$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\ dx$$

$$P(|X - a| < b) = P(-b < X - a < b)$$

- The mode of a c.r.v is the value of x which gives the maximum probability, i.e the x coordinate of the highest point in the domain.
- The **cumulative distribution function** (CDF):

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\ dt$$

$$\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to \infty} F(x) = 1$$

$$P(a < X < b) = F(b) - F(a)$$

$$\frac{d}{dx}F(x) = f(x)$$

- $F(x)$ is continuous and increasing (since $f(x) > 0$).
- To find the median m, set $F(m) = \frac{1}{2}$ and solve for m, i.e: $\int_{-\infty}^{m} f(t)dt = 0.5$

5.5 The Normal distribution

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

- The Normal distribution is a bell curve symmetrical about $x = \mu$.
- The mean = median = mode = μ.
- μ affects the location of the curve, whereas σ^2 affects the spread.
- The standard normal distribution is denoted by $Z \sim N(0, 1)$.
- Any normal distribution can be standardised: $Z = \frac{X-\mu}{\sigma}$
- The Z score represents the number of standard deviations away from the mean.
- To find c given $P(X < c) = p$, use invNorm.
- If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, then $aX + bY$ also has a normal distribution.

$$\begin{aligned} E(aX + bY) &= aE(X) + bE(Y) \\ &= a\mu_1 + b\mu_2 \\ \text{Var}(aX + bY) &= a^2\sigma_1^2 + b^2\sigma_2^2 \end{aligned}$$

$$aX + bY \sim N(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2)$$

5.6 Sampling

- If X is a random variable, $X_1, X_2, X_3, ..., X_n$ are a sample of n independent observations.
- The sample mean:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{nE(X)}{n} = E(X) = \mu$$

$$\text{Var}\big(\bar{X}\big) = \text{Var}\bigg(\frac{X_1 + X_2 + ... + X_n}{n}\bigg) = \frac{1}{n^2}\text{Var}(X_1 + X_2 + ... + X_n) = \frac{n\text{Var}(X)}{n^2} = \frac{\sigma^2}{n}$$

- For the sample sum: $E(S) = n\mu$, $\text{Var}(S) = n\sigma^2$
- Therefore, in a normal population:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \sum_{r=1}^{n} X_r \sim N(n\mu, n\sigma^2)$$

- The **Central Limit Theorem** states that, for a large sample size ($n \geq 50$), the sample mean/sum of a sample from *any* distribution (e.g not normal), will approximately follow the normal distribution.

5.7 Estimators

- An **estimator** is a test statistic T based on observed data that estimates an unknown parameter θ.
- The estimator is **unbiased** if $E(T) = \theta$.
- The sample mean is an unbiased estimator of μ since $E(\bar{X}) = \mu$.
- However, the sample variance is not an unbiased estimator for σ^2 since $E(S_n^2) = \frac{n-1}{n}\sigma^2$.
- An unbiased estimator for σ^2:

$$\begin{aligned} s_{n-1}^2 = \frac{n}{n-1} \times S_n^2 &= \frac{n}{n-1}\left(\frac{1}{n}\sum x^2 - (\bar{x})^2\right) \\ &= \frac{1}{n-1}\left(\sum x \;-\; \frac{(\sum x)^2}{n}\right) \end{aligned}$$

- An unbiased estimator is more **efficient** than another if it has a lower variance.

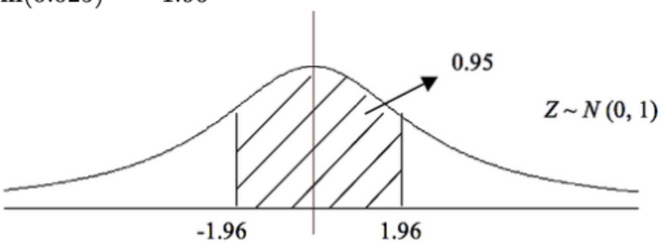
5.8 Confidence intervals

- A 95% confidence interval (CI) means that there is a 95% chance that the interval includes μ.
- For $X \sim N(\mu, \sigma^2)$, if we take a sample: $\bar{X} \sim N(\mu, \sigma^2)$.

$$\text{Confidence limits} = \bar{X} \pm Z_k \frac{\sigma}{\sqrt{n}}$$

$$\text{CI} = \left[\bar{X} - Z_k \frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_k \frac{\sigma}{\sqrt{n}}\right]$$

- Z_k is the **critical value**, and is found using invNorm.
- For a 95% CI: invNorm(0.025) $= -1.96$



- The width of a CI is $2Z_k \frac{\sigma}{\sqrt{n}}$
- If we have a large sample from any population (μ and σ^2 unknown), we can use the CLT.

$$\text{CI} = \left[\bar{x} - Z_k \frac{s_{n-1}}{\sqrt{n}},\ \bar{x} + Z_k \frac{s_{n-1}}{\sqrt{n}}\right]$$

- If the population is normal but we do not know the variance, we use the t-distribution.
  $T = \dfrac{\bar{X} - \mu}{s_{n-1}/\sqrt{n}}$ follows a t-distribution with $n-1$ degrees of freedom.

$$\text{CI} = \left[\bar{x} - t_k \frac{s_{n-1}}{\sqrt{n}},\ \bar{x} + t_k \frac{s_{n-1}}{\sqrt{n}}\right]$$

σ^2	n	Assumptions	Test Statistic
known	large	CLT	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
	small	normal	
unknown	large	CLT	$Z = \frac{\bar{X}-\mu}{s_{n-1}/\sqrt{n}} \sim N(0,1)$
	small	normal	$T = \frac{\bar{X}-\mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$

5.9 Hypothesis testing

1. State H_0 and H_1.
2. Test statistic.
3. Level of significance and rejection criteria.
4. Compute p-value (or z-value or t-value).
5. Conclusion in context.

e.g

$H_0 : \mu = 3$
$H_1 : \mu > 3$
Test statistic: $Z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
Sig level $= 5\%$, one tailed.
Reject H_0 if $p < 0.05$
Since p-value $= 0.03 < 0.05,$ we reject H_0 and conclude that there is significant evidence at the 5% level that...

- $P(\text{Type I Error}) = P(H_0 \text{ rejected}|H_0 \text{ true}) = \alpha\%$. i.e $P(\text{Type I Error})$ = **significance level**.
- $P(\text{Type II Error}) = P(H_0 \text{ accepted}|H_1 \text{ true})$.
- For example, for $H_0 : \mu = \mu_0 \;\; H_1 : \mu = \mu_1$,

$$P(\text{Type II Error}) = P(H_0 \text{ accepted}|H_1 \text{ true}) = P(\bar{X} < \text{critical value } |\bar{X} \sim N(\mu_1, \sigma^2))$$

5.10 PGFs

$$G(t) = E(t^X) = \sum t^x P(X = x)$$
$$G(1) = 1$$
$$G'(t) = \sum x t^{x-1} P(X = x) \therefore E(X) = G'(1)$$

$$G''(t) = \sum x(x-1) t^{x-2} P(X = x)$$
$$G''(1) = \sum x^2 P(X = x) - \sum x P(X = x) = E(X^2) - E(X)$$
$$\therefore E(X^2) = G''(1) + G'(1)$$
$$\therefore \mathrm{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$$

$$\text{If } Z = X + Y,\ G_Z(t) = E(t^Z) = E(t^{X+Y}) = E(t^X)E(t^Y) = G_X(t)G_Y(t)$$

- To find $P(X = n)$, we use the Maclaurin series: $P(X = n) = \frac{G^{(n)}(0)}{n!}$.
- To prove most things about PGFs, differentiation will be involved (sometimes using the product rule and chain rule).

Binomial

If $Y \sim B(n, p)$, we can say that $Y = X_1 + X_2 + X_3 + ... + X_n$ where X is a Bernoulli trial.

x	0	1
$P(X = x)$	q	p

$G_X(t) = \sum t^x P(X = x) = q + pt$

$G_Y(t) = E(t^Y) = E(t^{X_1 + ... + X_n}) = \left[E(t^X)\right]^n = [G_X(t)]^n = (q + pt)^n$

Poisson

If $X \sim Po(\lambda)$, $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

$$\begin{aligned} G(t) = E(t^X) &= \sum t^x P(X = x) \\ &= \sum t^x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum \frac{(\lambda t)^x}{x!} = e^{-\lambda} e^{\lambda t} = e^{\lambda(t-1)}. \end{aligned}$$

Geometric

If $X \sim Geo(p)$, $P(X = x) = pq^{x-1}$.

$$\begin{aligned} G(t) = E(t^X) &= \sum t^x P(X = x) \\ &= \sum t^x p q^{x-1} \\ &= pt + pt^2 q + pt^3 q^2 + pt^4 q^3 + ... + pt^n q^{n-1} + ... \\ S_\infty &= \frac{a}{1-r} = \frac{pt}{1-qt} \end{aligned}$$

Negative Binomial

If $Y \sim NB(r, p)$, we can say that $Y = X_1 + X_2 + X_3 + ... + X_r$, where $X \sim Geo(p)$.

$$G_Y(t) = E(t^Y) = E(t^{X_1+...+X_r}) = \left[E(t^X)\right]^r = [G_X(t)]^r = \left(\frac{pt}{1-qt}\right)^r$$

5.11 Bivariate data and correlations

- If X and Y are random variables, the joint probability distribution is $P(X = x \cap Y = y)$.
- $\sum\sum p(x, y) = 1$
- $E(XY) = \sum\sum xy\ p(x, y)$
- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$. X and Y independent $\implies \text{Cov}(X, Y) = 0$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$.
- The correlation coefficient measures the linear relationship between X and Y
$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$
- A **bivariate sample** consists of pairs of data (x_1, y_1). For a bivariate sample, the above points do not apply.
$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \text{ where } S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$
- If $r = 0$, there is no linear relationship, but it does not imply that X and Y are independent.
- r is independent of the units, and does not show any causality.
- In maths, **controlled variable = independent variable**.
- The y-on-x regression line $y = a + bx$ will always pass through $(\bar{x}, \bar{y})$.
$$y - \bar{y} = b(x - \bar{x}), \text{ where } b = \frac{S_{xy}}{S_{xx}}$$
- The x-on-y regression line is denoted by $x = c + dy$.
$$bd = r^2 \qquad r = \pm\sqrt{bd}, \text{ the sign depends on whether the gradient is positive or negative.}$$
- We can statistically test evidence of a correlation by assuming both variables follow a bivariate normal distribution with correlation coefficient ρ:
  $H_0 : \rho = 0$
  $H_1 : \rho \neq 0$
  Test statistic: $T = r\sqrt{\dfrac{n-2}{1-r^2}} \sim t_{n-2}$
  Sig level = 5%, two tailed.
  Reject H_0 if $|T| > \text{invt}(0.975, n-2)$
  Note: $T = r\sqrt{\dfrac{n-2}{1-r^2}}$ (sub in values)
  Since $|T| = 0.08 > \text{invt}(0.975, n-2)$, we reject H_0 and conclude that there is significant evidence at the 5% level that there is a correlation between...