1 Probability

Probability is the mathematical tool used to quantify the uncertainty in statistical inference. One must have a strong base in probability to start doing statistics.

Def: *Sample Space* is the collection of all possible outcomes of an experiment or process, denoted Ω

Def: An *Event* is a collection of outcomes in Ω, a subset of Ω, denoted $A, B, etc.$

Example: Tossing two coins: $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. Event A could be the event of one or more that one head occurring. $A = (H,H),(H,T),(T,H)$. Event B could be event of two tails occurring. $B = (T,T)$

1.1 Axioms of Probability

1. $P(\Omega) = 1$
2. $P(A) \geq$ for any event A
3. If the set of events A are mutually exclusive to each other then $P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} P(A_k)$

1.2 Properties of Probability

1. $P(A^c) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. If $A \subset B$, then $P(A) \leq P(B)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

1.3 Conditional Probability

Def: The probability of an event A given another event B.

- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A|B) \cdot P(B)$
- $P(B|A) = \dfrac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$

1.4 Independence

Def: The events A and B are independent iff the occurrence of one has nothing to do with the occurrence of the other.

- $P(A|B) = P(A)$ and $P(B|A) = P(B)$
- If A, B are independent, then $P(A \cap B) = P(A) \cdot P(B)$

2 Counting, Permutation & Combination

2.1 Counting

The number of ways to count collections of items depends on if we are removing items permanently or removing them, then putting them back in the lot of items.

Imagine we have n items, and we want to select k of them. If we make a selection, then place the selected item back in the group of items, the number of possible choices is given by

$$n * n * n * \cdots * n = n^k$$

Now if we make the choices, but then leave the choices out of the group, the number of possible ways to make r selections is given by

$$P_{k,n} = n * (n-1) * (n-2) * \cdots * (n-k-1) = \frac{n!}{(n-k)!}$$

Now let's do this again, assuming that the order of selection does not matter (we are still not replacing the items after they are selected). We must divide by $k!$ because there are that many possible orderings of the k items drawn.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

2.2 Permutation and Combination

Def: A *permutation* is an ordered arrangement of a set of objects. If we have n elements, and want to select k from them, then the permutation is given by $P_{k,n}$

Def: A *combination* is an unordered set of objects. If we have n elements, and want to select k from them, then the permutation is given by n choose $k = \binom{n}{k}$

Recall: $(x+y) = \sum_{k=0}^{n} \binom{n}{k} a^k \cdot b^{n-k}$

3 Discrete Distributions

3.1 Random Variables

Def: A *random variable* is a numerical valued function of a sample space Ω. These can be discrete or continuous depending on the sample space. We are mapping outcomes to numbers

- Let X be a random variable
- The event of $X = x$ stands for $\{\omega \in \Omega : W(\omega) = x\}$
- The event of $a \leq X \leq b$ stands for $\{\omega \in \Omega : a \leq X(\omega) \leq b\}$
- The event of $X \geq b$ stands for $\{\omega \in \Omega : X(\omega) \geq b\}$

3.2 Discrete Random Variables

Def: A *discrete random variable* is a random variable in a sample space with discrete and enumerable outcomes.

Def: A *probability distribution* of a discrete r.v. is a list of the distinct values x of the r.v. X, together with the associated probability. This is also called the *probability mass function* or *pmf*, and is given by the function

$$p(X) = P(X = x)$$

.

Def: A *cumulative distribution function* or *cdf* of a discrete r.v. is given by the function

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} p(x_i)$$

3.3 Bernoulli Distribution

Def: A *Bernoulli random variable* is a discrete r.v. whose only possible values are 0 and 1. Denoted $X \sim Ber(p)$ where p is the probability of a success. Properties:

- $P(X = 1) = p$ and $P(X = 0) = 1 - p$
- $E[X] = p$
- $\text{Var}(X) = np$
- $M(t) = (1 - p) + pe^(t)$
- $M^{(n)}(t) = pe^(t)$

Bernoulli random variables can be used as indicator functions. That is, for the event A, let 1_A denote an indicator function s.t. $1_A(\omega) = 1$ when $w \in A$ and 0 otherwise. Then $p = P(A)$.

3.4 Binomial Distribution

Def: A *Binomial distribution* is created by letting an r.v. X be equal to the number of successes in n Bernoulli trials where each trail has probability of success p. More formally, let $Z_1, Z_2, \cdots, Z_n$ be a series of independent and identically distributed (i.i.d) Bernoulli trials. Then $X = \sum_{i=0}^{n} Z_i$. Denoted $X \sim Bin(n, p)$. Properties:

- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $E[X] = n \cdot E[Z_i] = np$
- $\text{Var}(X) = np(1-p)$
- $M(t) = \left[pe^t + (1-p)\right]^n$
- $M'(t) = n\left[pe^t + (1-p)\right]^{n-1} \cdot (pe^t)$

3.5 Geometric Distribution

Def: A *Geometric distribution* is created by letting an r.v. X be equal to the number of Bernoulli trials until the first success where each trail has probability of success p. Denoted $X \sim Geom(p)$. Properties:

- $P(X = k) = p(1-p)^{k-1}$
- $E[X] = \dfrac{1}{p}$
- $\text{Var}(X) = \dfrac{1-p}{p^2}$
- $F(X) = 1 - (1-p)^k$
- $M(t) = \dfrac{pe^t}{1-(1-p)e^t}$

Note: The geometric distribution fulfills the memoryless property. Let $X \sim Geom(p)$, and let $x, x_0 > 0$. Then the memoryless property is given by:

$$P(X \geq x + x_0 | X \geq x_0) = P(X \geq x)$$

This is clearly explained by coin flips. If you have flipped a coin five times and gotten heads each time, what is the probability that you get heads on the next flip? Still .5

3.6 Negative Binomial Distribution

Def: A *negative binomial distribution* is created by a series of i.i.d. Bernoulli trials $Z_i \sim Ber(p)$. X is defined as the number of trials before r successes. The number r must be fixed. Properties:

- $P(X = k) = \binom{(k-1)}{(r-1)} p^r (1-p)^{k-r}$
- $E(X) = \frac{pr}{1-p}$
- $\text{Var}(X) = \frac{pr}{(1-p)^2}$
- $M(t) = \left(\frac{1-p}{1-pe^t}\right)$

Note: A geometric distribution is a special case of the negative binomial distribution where $r = 1$

3.7 Poisson

Def: A *Poisson* distribution can be used to approximate a binomial distribution when n is very large and p is very small, thus np is very normal sized. Recall that np is the expected value for a binomial distribution. We let $\lambda = np$ and create an r.v. $X \sim Pois(\lambda)$. Properties:

- $P(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$
- $E[X] = \lambda$
- $\text{Var}(X) = \lambda$
- $M(t) = e^{\lambda(e^t - 1)}$

Example: Show that the pmf of a Poisson r.v is a valid pmf

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{(pull out constants)}$$

$$e^{-\lambda} \cdot \sum_{x=0}^{\infty} \frac{\lambda^k}{k!} \qquad \text{(note the similarity to the Taylor expansion)}$$

$$e^{-\lambda}e^{\lambda}$$

$$1$$

4 Continuous Distributions

4.1 Continuous Random Variables

Def: A *continuous random variable* is a random variable in a sample space in which all numbers in a certain continuous interval are possible.

Def: A *probability distribution* of a continuous r.v. is a function that maps an outcome to its respective probability. This is also called the *probability density function*, or *pdf*, and is given by the functions $f(x)$ where:

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

Properties of $f(x)$:

- $f(x) \geq 0$ for all x
- $f(x)$ is piece wise continuous
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Def: The *cumulative distribution* of a continuous random variable is given by:

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

Properties of continuous random variables:

- $P(X = c) = 0$
- $P(a < X < b) = F(b) - F(a)$
- $P(a < X < b) = P(a \leq X \leq b)$
- For all x where $F'(x)$ exists, $F'(x) = f(x)$

Def: Let F be a strictly increasing cdf. Let $p \in (0, 1)$. Then $F^{-1}(p)$ is called the p-th quantile. This is essentially finding the x such that $F(x) = p$. In other words, given a probability p, find me the x such that the probability of a continuous r.v. being less than x is exactly p.

- the .5 quantile is the *median* of F
- the .25 quantile is the *lower quartile* of F
- the .75 quantile is the *upper quartile* of F

4.2 Exponential Distribution

Def: A continuous r.v. X follows an exponential distribution with parameter λ if:

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F(x,\lambda) = \begin{cases} 0 & \text{when } x < 0 \\ 1 - e^{-\lambda x} & \text{when } x \geq 0 \end{cases}$$

Note: The exponential distribution fulfills the memoryless property. Let $X \sim Exp(\lambda)$, and let $x, x_0 > 0$. Then the memoryless property is given by:

$$P(X \geq x + x_0 | X \geq x_0) = P(X \geq x)$$

Proof of the memoryless property:

$$\begin{aligned} P(X \geq x + x_0 | X \geq x_0) &= \frac{P(X \geq x + x_0 \cap X \geq x_0)}{P(X \geq x_0)} \\ &= \frac{P(X \geq x + x_0)}{P(X \geq x_0)} \\ &= \frac{1 - F(x + x_0)}{1 - F(x_0)} \\ &= \frac{e^{-\lambda(x+x_0)}}{e^{-\lambda x_0}} \\ &= \frac{e^{-\lambda x} \cdot e^{-\lambda x_0}}{e^{-\lambda x_0}} \\ &= e^{-\lambda x} \\ &= 1 - F(x) \\ &= P(X \geq x) \end{aligned}$$

Properties of exponential random variable:

- $E[X] = \dfrac{1}{\lambda}$
- $\text{Var}(X) = \dfrac{1}{\lambda^2}$
- $M(t) = \frac{\lambda}{\lambda - t}$

4.3 Gamma Distribution

Def: A continuous r.v. X follows a gamma distribution ($X \sim \Gamma(\lambda, \alpha)$) with shape parameter α and scale parameter λ if:

$$f(x, \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{a-1} e^{-\lambda x} \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}$$

The Gamma function is defined as:

$$\Gamma(\alpha) = \int_0^\infty x^{a-1} e^{-x} dx$$

Properties of the Gamma function:

- for any $\alpha > 0$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
- $\Gamma(n) = (n-1)!$ for any $n \in \mathbb{N}$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
- $M(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha$
- $M'(0) = E(X) = \frac{\alpha}{\lambda}$
- $\text{Var}(X) = \frac{\alpha}{\lambda^2}$

Note: The exponential distribution is a special case of gamma distribution where $\alpha = 1$.

4.4 Beta Distribution

Def: A continuous r.v. X follows a beta distribution ($X \sim Beta(a, b)$) with parameters a, b if $0 < x < 1$ and:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} x^{a-1} (1-x)^{b-1}$$

Note: When $a = b = 1$, $X \sim Unif(0, 1)$ Properties of the beta distribution:

- $E(X) = \frac{a}{a+b}$
- $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$

4.5 Uniform Distribution

Def: A continuous r.v. X follows a uniform distribution ($X \sim Unif(a, b)$) if:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The cdf for a uniform distribution is given by:

$$f(x) = \begin{cases} \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Properties of the uniform distribution:

- $E(X) = \frac{1}{2}(a+b)$
- $\text{Var}(X) = \frac{1}{12}(b-a)^2$
- $M(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$ when $t \neq 0$. $M(t) = 1$ when $t = 0$.

Note: A particular use for uniform distributions involves this property: Let $U \sim [0,1]$ and $X = F^{-1}(U)$, then the cdf of X is F.

$$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(X)$$

4.6 Normal Distribution

Def: A continuous r.v. X follows a normal distribution ($X \sim N(\mu, \sigma^2)$) with mean of μ and std dev of σ^2 and:

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

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

Note: Let r.v. $X \sim N(0,1)$. We call X the standard normal distribution, and it is often denoted Z. The cdf of Z is given by:

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^2}$$

$$\Phi(z) = P(Z \leq z)$$

Often times we will standardize a normal distribution to make it easier to understand. Let us say we have $X \sim N(\mu, \sigma^2)$, we know that $\frac{X-\mu}{\sigma} \sim N(0,1) = Z$.

Properties of the normal distribution:

- $M(t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$

4.7 Quantiles of Normal Distribution

Def: We call $z_\alpha = 100(1-\alpha)$ the $(1-a)$ quantile, of the standard normal distribution. z_a is the value for which the α-area lies to the *right*.

Let $X \sim N(\mu, \sigma^2)$ and let η_p be the p-quantile of X. Now we standardize:

$$p = P(X \leq \eta_p) = \Phi\left(\frac{\eta_p - \mu}{\sigma}\right)$$

Thus now we have:

$$\frac{\eta_p - \mu}{\sigma} = z_{1-p}$$

Or, more usefully:

$$\eta_p = \mu + \sigma z_{1-p}$$

4.8 Functions of Random Variables

What do we do when we want to construct new pdfs by applying functions to other pdfs?

Example: Let $X \sim N(0,1)$ and $Y = X^2$. Find pdf of Y.

First we construct $F_Y(y)$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \end{aligned}$$

Now we use $F_Y'(y) = f_Y(y)$.

$$\begin{aligned} f_Y(y) &= F_Y'(y) \\ &= \frac{1}{2}\phi(\sqrt{y})y^{\frac{-1}{2}} - \frac{-1}{2}\phi(-\sqrt{y})y^{\frac{-1}{2}} && \text{(note that } \phi(-\sqrt{y}) = \phi(\sqrt{y})\text{)} \\ &= y^{\frac{-1}{2}}\phi(\sqrt{y}) && \text{(only for } y \geq 0\text{)} \end{aligned}$$

But there is another way to do this. Suppose we have r.v. X with pdf of f_X and cdf of F_X. Let $Y = g(X)$.

We want to find f_Y and F_Y. If g is a differentiable and strictly monotonic function on a certain interval, and $f_X(x) = 0$ outside of that interval, then:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right|$$

Assume we have some distribution X, and with that we have the f_X and F_X. Now we apply some function g to X to yield a new distribution $Y = g(X)$. We want to find f_Y. We begin by obtaining F_Y.

$$\begin{aligned} F_Y(y) &= P(Y < y) \\ &= P(g(X) < y) \\ &= P(X < g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

Now that we have found F_Y we must differentiate it to find f_y.

$$\begin{aligned} f_Y(y) &= F'_Y(y) \\ &= \frac{d}{dx}\Big[F_X(g^{-1}(y))\Big] \\ &= F'_X(g^{-1}(y)) \cdot \frac{d}{dx}\Big[g^{-1}(y)\Big] \\ &= f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right| \end{aligned}$$

4.9 Summing Distributions

Suppose that X and Y are two discrete r.v.s have in the joint pmf of $p(x, y)$. Now we let the new r.v. $Z = X + Y$. Note that $Z = z$ exactly when $X = x$ and $Y = z - x$. So in order to find $p_Z(z)$, we must sum up all values:

$$p_Z(z) = \sum_{x=-\infty}^{\infty} p(x, z - x)$$

And if X and Y are independent of each other, we have:

$$p_Z(z) = \sum p_X(x) \cdot p_Y(z - x)$$

This is called the *convolution* of P_X and P_Y.

For the continuous case we a have similar formula:

$$f_Z(z) = \int_{-\infty}^{\infty} p(x, z - x)dx$$

Or if they are continuous then we have:

$$f_Z(z) = \int p_X(x) \cdot p_Y(z - x)dx$$

5 Joint Distributions

5.1 Discrete Random Variables

Def: The *joint probability mass function*, or *pmf*, or sometimes even *pdf*, for a pair of discrete r.v.s is given by:

$$p(x, y) = P(X = x \text{ and } Y = y)$$

Note: The joint pmf must satisfy:

- $p(x, y) \geq 0$ for all (x, y)
- $\sum_x \sum_y p(x, y) = 1$

Def: The *marginal pmf* of X is $p_X(x) = \sum_y p(x, y)$. Similarly, the marginal pmf of Y is $p_Y(y) = \sum_x p(x, y)$.

Def: Two r.v.s are *independent*, if and only if for every pair (x, y), we have $p(x, y) = p_X(x) \cdot p_Y(y)$

Def: The *joint cumulative distribution function*, or *cdf*, of two r.v.s is given by:

$$F(x, y) = P(X \leq x \text{ and } Y \leq y) = \sum_{x_i \leq x, y_i \leq y} p(x, y)$$

5.2 Categorical Distribution

Def: The *categorical distribution* is the generalized version of the Bernoulli distribution. Recall that a Bernoulli distribution is a test that can either succeed or fail. a categorical distribution can be any number of r outcomes, where each of the r outcomes occurs with a probability p_r s.t. $\sum p_r = 1$.

5.3 Multinomial Distribution

Def: The *multinomial distribution* is the generalized version of the binomial distribution. It involves a series of n categorical distributions where each categorical distribution has r possible outcomes. When $\sum n_i \neq n$ the probability is 0, because the number of outcomes that occur must sum to the number of trials. When $\sum n_i = n$ its pmf is given by:

$$p(n_1, n_2, \ldots, n_r) = \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

Note that:

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

5.4 Continuous Random Variables

Def: If X and Y are two continuous r.v.s, then $f(x,y)$ is the *join probability density function*, or *pdf*, of (X,Y) for any two dimensional set A:

$$P((X,Y) \in A) = \iint_A f(x,y)dxdy$$

Note: The joint pdf must satisfy:

- $f(x,y) \geq 0$ for all (x,y)
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) = 1$

Def: The *marginal pdf*s of X and Y are given by:

$$f_X(x) = \int_{\infty}^{\infty} f(x,y)dy \qquad f_Y(y) = \int_{\infty}^{\infty} f(x,y)dx$$

Def: Two continuous r.v.s are *independent* if for every pair (x,y) we have $f(x,y) = f_X(x) \cdot f_Y(y)$

Def: The *joint cumulative distribution function*, or *cdf*, of two r.v.s X and Y is given by:

$$F(x,y) = P(X < x \text{ and } Y < y) = \int_{-\infty}^{x}\int_{\infty}^{y} f(x,y)dydx$$

5.4.1 Joint Uniform Distribution

Def: This distribution occurs when the probability is evenly some area A. In this case $f(x,y) = \frac{1}{A}$. This must be the case so that integrating over the entire area equals one.

5.5 Transformations on Joint Distributions

Consider $(Y_1, Y_2) = g(X_1, X_2) = (g_1(X_1, X_2), g_2(X_1, X_2))$ where g is invertable and differentiatable. The joint pdf of Y_1 and Y_2 is given by:

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(h_1(y_1,y_2), h_2(y_1,y_2)) \left| \frac{1}{J(h_1(y_1,y_2), h_2(y_1,y_2))} \right|$$

Recall that:

$$J(x_1,x_2) = \det \begin{bmatrix} \frac{\delta g_1}{x_1} & \frac{\delta g_1}{x_2} \\ \frac{\delta g_2}{x_1} & \frac{\delta g_2}{x_2} \end{bmatrix}$$

6 Conditional Distributions

6.1 Discrete Random Variables

Def: For two discrete r.v.s X and Y, the *conditional pmf of X given Y* is:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Note: Rearranging the equation yields $p_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y)$, which brings us to an important property that:

$$P_X(x) = \sum_y P_{X|Y}(x|y)P_Y(y)$$

6.2 Continuous Random Variables

Def: For two continuous r.v.s X and Y, the *conditional pdf of X given Y* is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Note: Rearranging the equation yields $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$, which brings us to an important property that:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy$$

7 Expectation

Def: For a discrete r.v. X, we define *expectation* to be

$$E(X) = \sum_x x_i \cdot p(x_i)$$

And then for a continuous r.v. we define it to be

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

This value is also known as the *mean* of a r.v. and may be denotes μ or μ_X. Note that if X and Y are independent, then $E(XY) = E(X)E(Y)$. Also we have that $E(a + bX) = b \cdot E(X)$.

7.1 Expectation of Functions of r.v.s

Def: If we have a r.v. X and then another r.v. $Y = g(X)$ then we have:

$$E(Y) = \sum_x g(x_i)p(x_i) \text{ or } \int g(x)f(x)\,dx$$

7.2 Conditional Expectation

Def: The conditional expectation of Y given $X = x$ is:

$$E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x) \text{ or } \int y \cdot f_{Y|X}(y|x)$$

7.3 Properties of Expectation

Note: An important property of expectation is that:

$$E(Y) = E\big[E(Y|X)\big]$$

This yields something called the law of total expectation which is given by:

$$E(Y) = \sum_x E(Y|X=x)p_X(x) \text{ or } \int E(Y|X=x)f_X(x)\,dx$$

8 Variance

Def: The variance of a random variable X is given as:

$$\text{Var}(X) = E\Big\{\big[X - E(X)\big]^2\Big\} = E(X^2) - \big[E(X)\big]^2$$

Properties of variance:

- $\text{Var}(a + bX) = b^2 \cdot \text{Var}(X)$
- $\text{Var}(Y) = \text{Var}\big[E(Y|X)\big] + E\big[\text{Var}(Y|X)\big]$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- $\text{Var}\big(\sum(X_i)\big) = \sum \text{Var}(X_i)$ where X_i are independent

8.1 Covariance

Def: Covariance of two r.v.s is given by:

$$\text{Cov}(X, Y) = E\big[(X - \mu_X)(Y - \mu_Y)\big]$$

Properties of covariance:

- $\text{Cov}(aW + bX, cY + dZ) = ac\text{Cov}(W, Y) + bc\text{Cov}(X, Y) + bd\text{Cov}(X, Z) + ad\text{Cov}(W, Z)$
- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
- $\text{Cov}(X, X) = \text{Var}(X)$
- X, Y independent, then $\text{Cov}(X, Y) = 0$

Def: Now we define the correlation coefficient to be:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

8.2 Conditional Variance

Def: The conditional variance of Y given $X = x$ is:

$$\text{Var}(Y|X=x) = E\Big[\big\{Y - E(Y|X=x)\big\}^2\Big|X=x\Big] = E(Y^2|X=x) - E(Y|X=x)^2$$

9 Moment Generating Function

Def: The moment generating function for an r.v. X is given as $M(t) = E(e^{tX})$ thus we have:

$$M(t) = \sum_x e^{tX} p(x) = \int e^{tX} f(x)\, dx$$

Note: These are important because of the property that $M^{(r)}(0) = E(X^r)$

Theorem: Another important property of the moment generating function is that if X has mgf of $M_X(t)$ and $Y = a + bX$ then $M_Y(t) = e^{at} M_X(bt)$.

Theorem: If X and Y are independent functions and $Z = X + Y$, then we have $M_Z(t) = M_X(t) M_Y(t)$. This is due to expectation for two independent r.v.s

10 Distributions Derived from Normal

10.1 Chi Squared Distribution

Def: The Chi Squared distribution with 1 degree of freedom is given by U, where $U = Z^2$. Then the Chi Squared distribution with n degrees of freedom is given by $V = U_1 + U_2 + \cdots + U_n$ and is denoted X_n^2. Note that $V \sim Gamma(\frac{n}{2}, \frac{1}{2})$. The pdf is given:

$$f(v) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} v^{\frac{n}{2}-1} e^{\frac{-v}{2}}$$

Properties of the Chi Squared:

- $E(X) = n$
- $\text{Var}(X) = 2n$
- $M(t) = (1 - 2t)^{\frac{-n}{2}}$

11 Maximum Likelihood Estimator

Def: Suppose we have random variables $X_1, X_2, \dot{;} X_n$ with a joint density of $f(x_1, x_2, \ldots, x_n|\theta)$. Given observed values $X_i = x_i$, then the likelihood of θ as a function of $x_1, x_2, \ldots, x_n$ is defined as:

$$\text{Lik}(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$$

The *maximum likelihood estimate* of θ is the value of θ that maximizes the likelihood of the observed data. If the X_is are independent then we have:

$$\text{Lik}(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

Usually we try to maximize the log of the likelihood so we have:

$$l(\theta) = \sum_{i=1}^{n} \log\left[f(X_i|\theta)\right]$$

Then we must find:

$$S(\theta) = \frac{\delta}{\delta\theta} l(\theta)$$

Then solve for $\hat{\theta}$ such that $S(\hat{\theta}) = 0$ and check that $\hat{\theta}$ maximizes $l(\theta)$. Then $\hat{\theta}$ is the MLE.