Probability And Statistics

Mathematical Methods

Log and Exponential

log = $\log_e$ = ln; log(xy) = log(x) + log(y); log(x^y) = ylog(x); log(e^x) = x; $\lim_{x\to 0}$log(x) = -inf

e^x = exp(x); exp(x+y) = exp(x)exp(y); exp(x)^y = exp(xy); exp(log(x)) = x; exp(0) = 1

Arithmetic Progressions n^{th} term = a + (n-1)d; S_n = n/2(2a + (n-1)d); S_{inf} = +- inf (unless a = d = 0).

Geometric Progressions n^{th} term = ar^{n-1}; S_n = $(a(1-r^n))/(1-r)$ where r!=1; S_{inf} = a/(1-r) where |r| < 1.

Calculus (Differentiation opposite of Integration)

$$\frac{df}{dx} \equiv f'(x) \equiv \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}.$$

Fundamental Theorem of Calculus:

$$\frac{d}{dx}\int_{u=a}^{x} f(u)du = f(x),$$

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

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

Change of variable: if $y = g(x)$, $\int_a^b f(x)dx = \int_{g(a)}^{g(b)} f(g^{-1}(y))g^{-1\prime}(y)dy$

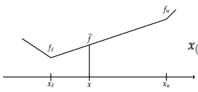
Quotient Rule: for $g(x) \neq 0$, $\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}$

By parts: $\int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx$

Images and inverses f: X→Y, A is subset of X (Image, Inverse, Inverse Image (B subset of Y)).

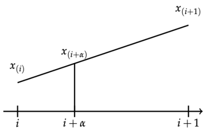
$f(A) = \{y \in Y | f(x) = y \text{ for some } x \in A\}$. $f^{-1}(f(x)) = x$ $f^{-1}(B) = \{x \in X | f(x) \in B\}$.

Interpolation



$$x_{(i+\alpha)} = (1-\alpha)\, x_{(i)} + \alpha\, x_{(i+1)}$$

$$\hat{f}(x) = f_{x_l} + (x - x_l)\frac{(f_{x_u} - f_{x_l})}{(x_u - x_l)}.$$



Numerical Summaries

Measures of Location Mean, Order statistic x_i i^{th} smallest value, median, mode most frequent – multimodal.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad x_{(n)} \equiv \max(x_1,\ldots,x_n) \quad x_{(1)} \equiv \min(x_1,\ldots,x_n) \quad \text{median} = x_{(\{n+1\}/2)} = \begin{cases} x_{(\{n+1\}/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even}\end{cases}$$

Others Geometric mean (less affected by large values), Harmonic mean (useful when averaging rates).

$$x_G = \sqrt[n]{\prod_{i=1}^{n} x_i}.\ x_G = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log x_i\right\} \quad x_H = \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right\}^{-1}$$

$x_A >= x_G >= x_H$

Measures of Dispersion Range, 1^{st} and 3^{rd} quartile,

first quartile $= x_{(\{n+1\}/4)}$

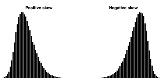
range $= x_{(n)} - x_{(1)}$ third quartile $= x_{(3\{n+1\}/4)}$ interquartile range = third quartile – first quartile

Five-point summary of set of data lists, in order min, lower quartile, sample median, upper quartile, max.

Sample variance (mean square)/Sample standard deviation (root mean square), Skewness (asymmetry)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad s^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \quad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

	Least Robust	More Robust	Most Robust
Location	$\frac{x_{(1)}+x_{(n)}}{2}$	$\bar{x}$	$x_{(\{n+1\}/2)}$
Dispersion	$x_{(n)} - x_{(1)}$	s^2	$x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}$



Covariance and correlation

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \quad s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}. \quad r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n s_x s_y}.$$

Box-and-Whisker Plots Median, 3^{rd}, and 1^{st} quartiles lines; whiskers extend to points within 3/2 IQR, extreme points plotted individually with x or o.

Empirical CDF Returns proportion of data having values which do not exceed x. $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$

Elementary Set Theory

Sets and notation (below), complement (of A is set of elems not in A), subset (some elems), singleton (1).

$\in$	-	"is an element of" (set membership)
$\iff$	-	"if and only if" (equivalence)
$\implies$	-	"implies"
$\exists$	-	"there exists"
$\forall$	-	"for all"
s.t. *or* |	-	"such that"
wrt	-	"with respect to"

COMMUTATIVITY	$A\cup B = B\cup A$ $A\cap B = B\cap A$
ASSOCIATIVITY	$A\cup(B\cup C) = (A\cup B)\cup C$ $A\cap(B\cap C) = (A\cap B)\cap C$
DISTRIBUTIVITY	$A\cup(B\cap C) = (A\cup B)\cap(A\cup C)$ $A\cap(B\cup C) = (A\cap B)\cup(A\cap C)$
DE MORGAN'S LAWS	$\overline{(A\cup B)} = \bar{A}\cap\bar{B}$ $\overline{(A\cap B)} = \bar{A}\cup\bar{B}$

Disjoint (A∩B = ϕ <- Null set); ***Partition of B*** Disjoint sets whose union form B; ***Difference*** A\B = A∩not B.

Cartesian Product Set of all ordered pairs of elements of two sets. ***Cardinality*** No. of elems in set.

|AUB| = |A| + |B| - |A∩B|

Ω	universal set
$\emptyset$	empty set
$A \subseteq \Omega$	Subset of Ω
$\bar{A}$	Complement of A
$\lvert A\rvert$	Cardinality (or size) of A
$A\cup B$	union (A or B)
$A\cap B$	intersection (A and B)
$A \subset B$	set inclusion (elements of A are also in B)
$A\backslash B$	set difference (elements in A that are not in B)

Probability

Sample spaces (S) Set of all possible outcomes of ***experiment***.

Event Subset of sample space (extreme events ϕ (the null event) or S). E occurs if outcome s∈E.

Elementary events Singleton subsets of S which contain exactly one element from S.

Mutually Exclusive events If they are disjoint, i.e., at most one event can occur.

Sigma Algebra F is subcollection of sets of all subsets of S with the following properties:

a) if $E, F \in \mathcal{F}$ then $E\cup F \in \mathcal{F}$ and $E\cap F \in \mathcal{F}$;

b) if $E \in \mathcal{F}$ then $\bar{E} \in \mathcal{F}$;

c) $\emptyset \in \mathcal{F}$.

Definition 5.2.1. *A collection $\mathcal{F}$ of subsets of S is called a σ-**field** (or σ-**algebra**) if it satisfies the following conditions:*

a) $\emptyset \in \mathcal{F}$;

b) if $E_1, E_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$;

c) if $E \in \mathcal{F}$ then $\bar{E} \in \mathcal{F}$.

Probability Measure

Definition 5.3.1. *A **probability measure** P on $(S, \mathcal{F})$ is a mapping $\mathrm{P}: \mathcal{F} \to [0,1]$ satisfying*

a) $\mathrm{P}(S) = 1$;

b) if $E_1, E_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, so that $E_i \cap E_j = \emptyset$ for all pairs i, j with $i \neq j$, then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty}\mathrm{P}(E_i)$$

*The triple $(S, \mathcal{F}, \mathrm{P})$, consisting of a set S, a σ-algebra $\mathcal{F}$ and probability measure P on $(S, \mathcal{F})$ is called a **probability space**.*

Interpretations of Probability

Classical Equally likely (uniform) probabilities.

Frequentist Repeated observations and find limiting value.

Subjective Degree of belief held by individual.

Independent Events

E, F independent iff P(E∩F)=P(E)P(F).

Probability space denoted by triple (S, F, P) has the following properties for E, F ⊆ S:

1. $\mathrm{P}(\bar{E}) = 1 - \mathrm{P}(E)$
2. if $E \subseteq F$, then $\mathrm{P}(E) \leq \mathrm{P}(F)$.
3. In general, $\mathrm{P}(E\cup F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E\cap F)$.
4. $\mathrm{P}(E\cap\bar{F}) = \mathrm{P}(E) - \mathrm{P}(E\cap F)$.
5. $\mathrm{P}(E\cup F) \leq \mathrm{P}(E) + \mathrm{P}(F)$.
6. $\mathrm{P}(E\cap F) \geq \mathrm{P}(E) + \mathrm{P}(F) - 1$.

	⚀	⚁	⚂	⚃	⚄	⚅	
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
T	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Probability Table Can calculate probability of any event of interest (right side above).

Conditional Probability P(E|F) = P(E∩F)/P(F); ***Conditional Independence*** P(E_1∩E_2|F) = P(E_1|F)P(E_2|F).

Theorem of Total Probability Let E_1, …, E_k be partition on S, then for any event F ⊆ S we have P(F) = sum from i to k P(F|E_i)P(E_i). ***Bayes' Theorem*** Let E_1, …, E_k be partition on S P(E_i|F) = (P(F|E_i)P(E_i))/P(F) = (P(F|E_i)P(E_i))/(sum from 1 to k P(F|E_i)P(E_i)).

Conditional Probability P(E|F); ***Joint Probability*** P(E∩F); ***Marginal Probability*** P(E).

Summary of Conditional Probability

1. If $\mathrm{P}(F) > 0$ then
$$\mathrm{P}(E|F) = \frac{\mathrm{P}(E\cap F)}{\mathrm{P}(F)}$$
2. $\mathrm{P}(\cdot|F)$ satisfies the axioms of probability, for fixed F. However, in general, $\mathrm{P}(E|\cdot)$ does not satisfy the axioms of probability, for fixed E.
3. In general, $\mathrm{P}(E|F) \neq \mathrm{P}(F|E)$.
4. E and F are independent if and only if $\mathrm{P}(E|F) = \mathrm{P}(E)$.

Discrete Random Variables

Random Variables Measurable mapping X: S → IR with property that {s∈S : X(s) <= x} ∈ F for each x ∈ IR.

Probability Distribution Function P_X(X <= x) = P(S_x); ***Range*** of random variable is image of S under X: $\mathbb{X}$ = X(S) = {x ∈ IR|∃s ∈ S s.t. X(s) = x}.

Cumulative Distributive Function F_X(x) = P_X(X<=x)); Conditions of valid CDF (right side below). ***P_X(a<X<=b) = F_X(b) – F_X(a).***

Definition 6.1.3. *The* **cumulative distribution function** (**CDF**) *of a random variable X is the function $F_X: \mathbb{R} \to [0,1]$, defined by*

$$F_X(x) = \mathrm{P}_X(X \leq x)$$

i) $0 \leq F_X(x) \leq 1, \forall x \in \mathbb{R}$;

ii) Monotonicity: $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;

iii) $F_X(-\infty) = 0, F_X(\infty) = 1$.

Discrete Random Variable X is discrete if range of X, is countable (finite or infinite).

Probability Mass Function (Probability Function) – For discrete random variable X, p_x(x) = P(X=x), x ∈ X. If x ∉ X, p_x(x) = 0. Properties of pmf (left), connection between F_x and p_x (right)

i) $p_X(x_i) \geq 0$;

ii) $\sum_{x\in\mathbb{X}} p_X(x) = 1$. $\quad F_X(x) = \sum_{x_i \leq x} p_X(x_i) \iff p_X(x_i) = F_X(x_i) - F_X(x_{i-1}), \quad i = 2, 3, \ldots.$

6.2.4 Properties of Discrete CDF F_X

i) In the limiting cases,
$$\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to\infty} F_X(x) = 1.$$

ii) F_X is continuous from the right on $\mathbb{R}$, that is, for $x \in \mathbb{R}$,
$$\lim_{h\to 0^+} F_X(x+h) = F_X(x)$$

iii) F_X is non-decreasing, that is,
$$a < b \implies F_X(a) \leq F_X(b).$$

iv) For $a < b$
$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

Mean and Variance

Expectation or mean of a discrete random variable is defined to be E_X(X) = sum of all x (xp_x(x)).

$$\mathrm{E}(g(X)) = \sum_x g(x)p_X(x) \quad \mathrm{E}(ag(X) + bh(X)) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(X))$$

(i) For a linear function, $g(X) = aX + b$ for constants, we have (from Theorem 6.4) that

$$\mathrm{E}(g(X)) = \sum_x (ax+b)p_X(x) = a\sum_x xp_X(x) + b\sum_x p_X(x)$$

and since $\sum_x xp_X(x) = \mathrm{E}(X)$ and $\sum_x p_X(x) = 1$ we have

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$

Definition 6.4.1. *Let X be a random variable. The* **variance** *of X, denoted by σ^2 or σ_X^2 or $Var_X(X)$ is defined by*

$$Var_X(X) = \mathrm{E}_X[\{X - \mathrm{E}_X(X)\}^2].$$

Definition 6.4.2. *The* **standard deviation** *of a random variable X, written $sd_X(X)$ (or sometimes σ_X), is the square root of the variance.*

$$sd_X(X) = \sqrt{Var_X(X)}.$$

Definition 6.4.3. *The* **skewness** (γ_1) *of a discrete random variable X is given by*

$$\gamma_1 = \frac{\mathrm{E}_X[\{X - \mathrm{E}_X(X)\}^3]}{sd_X(X)^3}.$$

6.4.1 Sums of Random Variables

Let $X_1, X_2, \ldots, X_n$ be n random variables, perhaps with different distributions and not necessarily independent.

Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of those variables, and $\frac{S_n}{n}$ be their average.

Then the mean of S_n is given by

$$\mathrm{E}(S_n) = \sum_{i=1}^{n}\mathrm{E}(X_i), \quad \mathrm{E}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n}\mathrm{E}(X_i)}{n}.$$

However, for the variance of S_n, only if $X_1, X_2, \ldots, X_n$ are **independent**, we have

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i), \quad \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n}\mathrm{Var}(X_i)}{n^2}.$$

So if $X_1, X_2, \ldots, X_n$ are independent and identically distributed with $\mathrm{E}(X_i) = \mu_X$ and $\mathrm{Var}(X_i) = \sigma_X^2$ we get

$$\mathrm{E}\left(\frac{S_n}{n}\right) = \mu_X, \quad \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma_X^2}{n}.$$

Geometric Distribution Potentially infinite sequence of independent Bernoulli(p) random variables

$$X = \min\{i | X_i = 1\}$$

to be the index of the first Bernoulli trial to result in a 1.

Example Tossing a coin, X is the number of tosses until the first head is obtained, $p = \frac{1}{2}$. ■

Then X is a random variable taking values in $\mathbb{Z}^+ = \{1, 2, \ldots\}$, and we say $X \sim \mathrm{Geometric}(p)$.

Clearly the pmf is given by

$$p_X(x) = p(1-p)^{x-1}, \quad x \in \mathbb{X} = \{1, 2, \ldots\}, \quad 0 \leq p \leq 1.$$

Notes

- The mean and variance are
$$\mu \equiv \mathrm{E}(X) = \frac{1}{p}, \quad \sigma^2 \equiv \mathrm{Var}(X) = \frac{1-p}{p^2}.$$
- The skewness is given by
$$\gamma_1 = \frac{2-p}{\sqrt{1-p}},$$
and so is always positive.

Alternative Formulation

If $X \sim \mathrm{Geometric}(p)$, let us consider $Y = X - 1$.

Then Y is a random variable taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$, and corresponds to the number of independent Bernoulli(p) trials *before* we obtain our first 1. (Some texts refer to *this* as the Geometric distribution.)

Note we have pmf

$$p_Y(y) = p(1-p)^y, \quad y = 0, 1, 2, \ldots,$$

and the mean becomes

$$\mu_Y \equiv \mathrm{E}_Y(Y) = \frac{1-p}{p}.$$

while the variance and skewness are unaffected by the shift.

6.5.5 Discrete Uniform Distribution

Let X be a random variable on $\{1, 2, \ldots, n\}$ with pmf

$$p_X(x) = \frac{1}{n}, \quad x \in \mathbb{X} = \{1, 2, \ldots, n\}.$$

Then X is said to follow a discrete uniform distribution and we write $X \sim \mathrm{U}(\{1, 2, \ldots, n\})$.

Note The mean and variance are

$$\mu \equiv \mathrm{E}(X) = \frac{n+1}{2}, \quad \sigma^2 \equiv \mathrm{Var}(X) = \frac{n^2-1}{12}.$$

and the skewness is clearly zero.

Bernoulli Distribution Experiment with only 2 outcomes, encoded as random variable X taking value 1, probability p, or 0, probability (1-p). e.g., coin toss 1x.

Then we say $X \sim \mathrm{Bernoulli}(p)$ and note the pmf to be

$$p_X(x) = p^x(1-p)^{1-x}, \quad x \in \mathbb{X} = \{0, 1\}, \quad 0 \leq p \leq 1$$

Note Using the formulae for mean and variance, it follows that

$$\mu \equiv \mathrm{E}(X) = p, \quad \sigma^2 \equiv \mathrm{Var}(X) = p(1-p).$$

Binomial Distribution n identical, independent Bernoulli(p) trials, e.g., tossing coin n times and measuring number of heads obtained .

Then X is a random variable taking values in $\{0, 1, 2, \ldots, n\}$, and we say $X \sim \mathrm{Binomial}(n, p)$.

From the Binomial Theorem we find the pmf to be

$$p_X(x) = \binom{n}{x}p^x(1-p)^{n-x}, \quad x \in \mathbb{X} = \{0, 1, 2, \ldots, n\}, \quad n \geq 1, \quad 0 \leq p \leq 1.$$

Notes

- To calculate the Binomial pmf we recall that $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ and $x! = \prod_{i=1}^{x} i$. (Note $0! = 1$.)
- It can be shown, either directly from the pmf or from the results for sums of random variables, that the mean and variance are
$$\mu \equiv \mathrm{E}(X) = np, \quad \sigma^2 \equiv \mathrm{Var}(X) = np(1-p).$$
- The skewness is given by
$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}.$$

Poisson Distribution Number of random events occurring per unit time/space.

Let X be a random variable on $\mathbb{N} = \{0, 1, 2, \ldots\}$ with pmf

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \mathbb{X} = \{0, 1, 2, \ldots\}, \quad \lambda > 0.$$

Then X is said to follow a Poisson distribution with *rate* parameter λ and we write $X \sim \mathrm{Poisson}(\lambda)$.

Notes

- Poisson random variables are concerned with the number of random events occurring per unit of time or space, when there is a constant underlying probability 'rate' of events occurring across this unit.
  - the number of minor car crashes per day in the U.K.;
  - the number of potholes in each mile of road;
  - the number of jobs which arrive at a database server per hour;
  - the number of particles emitted by a radioactive substance in a given time.
- An interesting property of the Poisson distribution is that it has equal mean and variance, namely
$$\mu \equiv \mathrm{E}(X) = \lambda, \quad \sigma^2 \equiv \mathrm{Var}(X) = \lambda.$$
- The skewness is given by
$$\gamma_1 = \frac{1}{\sqrt{\lambda}},$$
so is always positive but decreasing as λ increases.

Continuous Random Variables

Definition 7.0.1. *A random variable X is (absolutely)* **continuous** *if $\exists f_X: \mathbb{R} \to \mathbb{R}$ (measurable) such that*

$$\mathrm{P}_X(B) = \int_{x\in B} f_X(x)dx, \quad B \subseteq \mathbb{R},$$

in which case f_X is referred to as the **probability density function**, *or* **pdf**, *of X.*

7.0.2 Properties of Continuous F_X and f_X

By analogy with the discrete case, let $\mathbb{X}$ be the range of X, so that $\mathbb{X} = \{x : f_X(x) > 0\}$.

i) For the cdf of a continuous random variable,
$$\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to\infty} F_X(x) = 1.$$

ii) At values of x where F_X is differentiable
$$f_X(x) = \frac{d}{dt}F_X(t)\Big|_{t=x} \equiv F_X'(x).$$

iii) If X is continuous,
$$f_X(x) \neq \mathrm{P}(X = x) = \lim_{h\to 0^+}[\mathrm{P}(X \leq x) - \mathrm{P}(X \leq x-h)] = \lim_{h\to 0^+}[F_X(x) - F_X(x-h)] = 0$$

Warning! People usually forget, that $\mathrm{P}(X = x) = 0$ for all x, when X is a continuous random variable.

iv) The pdf $f_X(x)$ is not itself a probability, then unlike the pmf of a discrete random variable we do not require $f_X(x) \leq 1$.

v) For $a < b$,
$$\mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X < b) = F_X(b) - F_X(a).$$

vi) From Definition 7.0.1 it is clear that the pdf of a continuous random variable X completely characterises its distribution, so we often just specify f_X.

It follows that a function f_X is a pdf for a continuous random variable X if and only if

i) $f_X(x) \geq 0$,

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

This result follows direct from the definitions and properties of F_X.

Suppose we are interested in whether a continuous random variable X lies in an interval $(a, b]$. Well, $\mathrm{P}_X(a < X \leq b) = \mathrm{P}_X(X \leq b) - \mathrm{P}_X(X \leq a)$, which in terms of the cdf and pdf gives

$$\mathrm{P}_X(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx.$$

Definition 7.0.2. *The* **cumulative distribution function** *of* **CDF**, *F_X of a continuous random variable X is defined as*

$$F_X(x) = \mathrm{P}(X \leq x), \quad x \in \mathbb{R}.$$

Suppose that X is a continuous random variable X with pdf f_X and cdf F_X. Let $Y = g(X)$ be a function of X for some (measurable) function $g: \mathbb{R} \to \mathbb{R}$ s.t. g is continuous and strictly monotonic (so g^{-1} exists) . We call $Y = g(X)$ a transformation of X.

Suppose g is monotonic increasing. We can compute the pdf and cdf of $Y = g(X)$ as follows:

The cdf of Y is given by

$$F_Y(y) = \mathrm{P}_Y(Y \leq y) = \mathrm{P}_Y(g(X) \leq y) = \mathrm{P}_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

The pdf of Y is given by using the chain rule of differentiation:

$$f_Y(y) = F_Y'(y) = f_X\{g^{-1}(y)\}g^{-1\prime}(y)$$

Note $g^{-1\prime}(y) = \frac{d}{dy}g^{-1}(y)$ is positive since we assumed g was increasing.

If g monotonic decreasing, we have that

$$F_Y(y) = \mathrm{P}_Y(Y \leq y) = \mathrm{P}_Y(g(X) \leq y) = \mathrm{P}_X(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

By comparison with before, we would have

$$f_Y(y) = F_Y'(y) = -f_X\{g^{-1}(y)\}g^{-1\prime}(y)$$

with $g^{-1\prime}(y)$ always negative.

Therefore, for $Y = g(X)$ we have

$$f_Y(y) = f_X\{g^{-1}(y)\}|g^{-1\prime}(y)|. \tag{7.1}$$

Continuous Random Variables

Mean, Variance and Quantiles

Mean/Expectation

$$\mu_X \text{ or } E_X(X) = \int_{-\infty}^{\infty} x f_X(x)dx. \quad \mathrm{E}_X\{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x)dx. \quad \mathrm{E}(aX+b) = a\mathrm{E}(X)+b,$$

$$\mathrm{E}\{g(X)+h(X)\} = \mathrm{E}\{g(X)\} + \mathrm{E}\{h(X)\}.$$

Variance

$$\sigma_X^2 \text{ or } Var_X(X) = E\{(X-\mu_X)^2\} = \int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)dx. \quad \mathrm{Var}_X(X) = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu_X^2 = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2.$$

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X),$$

Quantiles

Definition 7.1.3. *For a (continuous) random variable X we define the* **α-quantile** $Q_X(\alpha)$, $0 \le \alpha \le 1$ *to satisfy* $\mathrm{P}(X \le Q_X(\alpha)) = \alpha$,

$$Q_X(\alpha) = F_X^{-1}(\alpha).$$

Continuous Uniform Distribution

Pdf, cdf, a = 0 and b = 1 is referred to as **Standard Uniform**

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise,} \end{cases} \quad F_X(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases}$$

- Suppose $X \sim \mathrm{U}(0,1)$, so $F_X(x) = x$, $0 \le x \le 1$. We wish to map the interval $(0,1)$ to the general interval (a,b), where $a < b \in \mathbb{R}$. So we define a new random variable $Y = a + (b-a)X$, so $a < Y < b$.

  We first observe that for any $y \in (a,b)$,

  $$Y \le y \iff a + (b-a)X \le y \iff X \le \frac{y-a}{b-a}.$$

  From this we find $Y \sim \mathrm{U}(a,b)$, since

  $$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}\left(X \le \frac{y-a}{b-a}\right) = F_X\left(\frac{y-a}{b-a}\right) = \frac{y-a}{b-a}.$$

- To find the mean of $X \sim \mathrm{U}(a,b)$,

  $$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_a^b x.\frac{1}{b-a}dx = \left[\frac{x^2}{2(b-a)}\right]_a^b$$

  $$= \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

Similarly we get $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{(b-a)^2}{12}$, so

$$\mu = \frac{a+b}{2}, \qquad \sigma^2 = \frac{(b-a)^2}{12}.$$

Exponential Distribution

Pdf, cdf, θ = 1/λ parameter of distribution – expectation,

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \ge 0, F_X(x) = 1 - e^{-\lambda x}, \quad x > 0. \quad \mathrm{E}(X) = \frac{1}{\lambda} = \theta, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

If $X \sim \mathrm{Exp}(\lambda)$, then, for all $x, t > 0$,

$$\mathrm{P}(X > x+t | X > t) = \frac{\mathrm{P}(X > x+t \cap X > t)}{\mathrm{P}(X > t)} = \frac{\mathrm{P}(X > x+t)}{\mathrm{P}(X > t)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = \mathrm{P}(X > x).$$

Thus, for all $x, t > 0$, $\mathrm{P}(X > x+t | X > t) = \mathrm{P}(X > x)$ — this is known as the **Lack of Memory Property**, and is unique to the exponential distribution amongst continuous distributions.

Normal Distribution

Pdf, cdf,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad F_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}dt.$$

Special Case: If $\mu = 0$ and $\sigma^2 = 1$, then X has a **standard** or **unit** normal distribution. The pdf of the standard normal distribution is written as $\phi(x)$ and simplifies to

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}.$$

Also, the cdf of the standard normal distribution is written as $\Phi(x)$. Again, for the cdf, we can only write

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

$$F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}\left(Y \le \frac{x-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right). \qquad \Phi(z) = 1 - \Phi(-z).$$

Central Limit Theorem

$$\lim_{n\to\infty}\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \sim \Phi. \quad \lim_{n\to\infty}\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \Phi, \quad \text{where} \quad \overline{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Or for large n

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \sum_{i=1}^n X_i \sim \mathrm{N}\left(n\mu, n\sigma^2\right)$$

Example Consider the most simple example, that $X_1, X_2, \ldots$ are i.i.d. Bernoulli(p) discrete random variables taking value 0 or 1.

Then the $\{X_i\}$ have mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. Then, by definition, we know that for any n we have

$$\sum_{i=1}^n X_i \sim \mathrm{Binomial}(n,p).$$

which has mean np and variance $np(1-p)$.

But now, by the Central Limit Theorem (CLT), we also have for large n that approximately

$$\sum_{i=1}^n X_i \sim \mathrm{N}\left(n\mu, n\sigma^2\right) \equiv \mathrm{N}(np, np(1-p)).$$

So for large n

$$\mathrm{Binomial}(n,p) \approx \mathrm{N}(np, np(1-p)).$$

Jointly Distributed Random Variables

Marginal Probability distributions $P_X(B) = P(X^{-1}(B))$. Joint probability density $P_{XY}(B_X, B_Y) = P\{X^{-1}(B_X) \cap Y^{-1}(B_Y)\}$.

Joint CDF; Recovering Marginal CDFs for X and Y; Joint PMF; Recovering Marginal PMFs p_X and p_Y

$$F_X(x) = F_{XY}(x,\infty) \qquad p_X(x) = \sum_y p_{XY}(x,y), \quad p_Y(y) = \sum_x p_{XY}(x,y).$$

$$F_{XY}(x,y) = \mathrm{P}_{XY}(X \le x, Y \le y). \; F_Y(y) = F_{XY}(\infty, y). \; p_{XY}(x,y) = \mathrm{P}_{XY}(X = x, Y = y)$$

Properties of valid CDF; Properties of Joint PMFs

1. $0 \le F_{XY}(x,y) \le 1, \forall x,y \in \mathbb{R}$;
2. Monotonicity: $\forall x_1, x_2, y_1, y_2 \in \mathbb{R}$,

   $x_1 < x_2 \Rightarrow F_{XY}(x_1,y_1) \le F_{XY}(x_2,y_1)$ and $y_1 < y_2 \Rightarrow F_{XY}(x_1,y_1) \le F_{XY}(x_1,y_2)$;
3. $\forall x,y \in \mathbb{R}$,

   $F_{XY}(x,-\infty) = 0, F_{XY}(-\infty,y) = 0$ and $F_{XY}(\infty,\infty) = 1$.

1. $0 \le p_{XY}(x,y) \le 1, \forall x,y \in \mathbb{R}$;
2. $\sum_y\sum_x p_{XY}(x,y) = 1$.

Conditional Probability

$$E_{Y|X}(Y|X=x) = \int_{y=-\infty}^{\infty} y\, f(y|x)dy.$$

Correlation

$$\rho_{XY} = Cor(X,Y) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Joint PDFs

$$\mathrm{P}_{XY}(B_{XY}) = \int_{(x,y)\in B_{XY}} f_{XY}(x,y)dxdy, \; F_{XY}(x,y) = \int_{t=-\infty}^{y}\int_{s=-\infty}^{x} f_{XY}(s,t)dsdt, \; f_{XY}(x,y) = \frac{\partial^2}{\partial x\partial y}F_{XY}(x,y)$$

Recovering marginal densities

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}F_{XY}(x,\infty)$$

$$= \frac{d}{dx}\int_{y=-\infty}^{\infty}\int_{s=-\infty}^{x} f_{XY}(s,y)dsdy. \; f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x,y)dy, \quad f_Y(y) = \int_{x=-\infty}^{\infty} f_{XY}(x,y)dx.$$

Covariance

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = \mathrm{E}_{XY}[(X-\mu_X)(Y-\mu_Y)]$$

Independence; Independence; Conditional probability; conditional pdf

$$\mathrm{P}_{XY}(B_X,B_Y) = \mathrm{P}_X(B_X)\mathrm{P}_Y(B_Y) \; f_{XY}(x,y) = f_X(x)f_Y(y) \; \mathrm{P}_{Y|X}(B_Y|B_X) = \frac{\mathrm{P}_{XY}(B_X,B_Y)}{\mathrm{P}_X(B_X)} \quad f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Expectation (x4); Conditional expectation

$$E_{XY}\{g(X,Y)\} = \sum_y\sum_x g(x,y)p_{XY}(x,y) \; E_{XY}\{g(X,Y)\} = \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} g(x,y)f_{XY}(x,y)dxdy \; \mathrm{E}_{XY}(XY) = \mathrm{E}_X(X)\mathrm{E}_Y(Y)$$

$$\mathrm{E}_{XY}\{g_1(X)+g_2(Y)\} = E_X\{g_1(X)\} + E_Y\{g_2(Y)\} \; E_{Y|X}(Y|X=x) = \sum_y y\, p(y|x)$$

Estimation

Statistic Function T = T(X_1, …, X_n) = T(X) and is itself a random variable.

Finding the MLE

In general, we have the following procedure to find MLEs.

1. Write down the likelihood function, $L(\theta)$ where

   $$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

   that is, the product of the n mass/density functions viewed as a function of θ.
2. Take the natural log of the likelihood, and collect terms involving θ.
3. Find the value of θ for which log-likelihood is maximised. This is typically done by finding $\widehat{\theta}$ that solves

   $$\frac{\partial}{\partial\theta}\ell(\widehat{\theta}) = \frac{\partial}{\partial\theta}\log(L(\widehat{\theta})) = 0$$
4. Check that the estimate $\widehat{\theta}$ obtained in step 3 corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\ell(\theta)$ wrt θ. If

   $$\frac{\partial^2}{\partial\theta^2}\ell(\widehat{\theta}) < 0$$

   at $\theta = \widehat{\theta}$, then $\widehat{\theta}$ is confirmed as the MLE of θ.

Confidence Interval

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

$$\frac{\overline{X}-\mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

where $s_{n-1} = \sqrt{\frac{\sum_{i=1}^n(X_i-\overline{X})^2}{n-1}}$ is the bias-corrected sample standard deviation, and t_ν is the Student's t-distribution with ν degrees of freedom.

Then it follows that an exact $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\overline{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{n}}, \overline{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{n}}\right]$$

where $t_{\nu,\alpha}$ is the α-quantile of t_ν.

Example Continuing the Binomial question... each of our Binomial$(10,p)$ samples X_i have pmf

$$p_X(x_i) = \binom{10}{x_i}p^{x_i}(1-p)^{10-x_i}, \qquad i = 1,2,\ldots,100.$$

Since the $n = 100$ data samples are assumed independent, the likelihood function for p for all of the data is

$$L(p|\underline{x}) = L(p) = \prod_{i=1}^n p_X(x_i) = \prod_{i=1}^n\left\{\binom{10}{x_i}p^{x_i}(1-p)^{10-x_i}\right\}$$

$$= \left\{\prod_{i=1}^n\binom{10}{x_i}\right\}p^{\sum_{i=1}^n x_i}(1-p)^{10n-\sum_{i=1}^n x_i}.$$

So the log-likelihood is given by

$$\ell(p) = \log\left\{\prod_{i=1}^n\binom{10}{x_i}\right\} + \log(p)\sum_{i=1}^n x_i + \log(1-p)\left(10n - \sum_{i=1}^n x_i\right).$$

Next, we differentiate $\ell(p)$

$$\frac{\partial}{\partial p}\ell(p) = 0 + \frac{\sum_{i=1}^n x_i}{p} - \frac{10n - \sum_{i=1}^n x_i}{1-p}.$$

Setting this derivative equal to zero, we get

$$\frac{\sum_{i=1}^n x_i}{\widehat{p}} - \frac{10n - \sum_{i=1}^n x_i}{1-\widehat{p}} = 0 \Rightarrow (1-\widehat{p})\sum_{i=1}^n x_i = \widehat{p}\left(10n - \sum_{i=1}^n x_i\right)$$

$$\Rightarrow \sum_{i=1}^n x_i = \widehat{p}\left(10n - \sum_{i=1}^n x_i + \sum_{i=1}^n x_i\right)$$

$$\Rightarrow \widehat{p} = \frac{\sum_{i=1}^n x_i}{10n} = \frac{\overline{x}}{10}.$$

To check this point is a maximum of ℓ, we find the second derivative

$$\frac{\partial^2}{\partial p^2}\ell(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{10n - \sum_{i=1}^n x_i}{(1-p)^2} = -\frac{n\overline{x}}{p^2} - \frac{10n - n\overline{x}}{(1-p)^2} = -n\left(\frac{\overline{x}}{p^2} + \frac{10-\overline{x}}{(1-p)^2}\right)$$

(which is in fact $< 0\ \forall p$, the likelihood is *log concave*).

Substituting $\widehat{p} = \frac{\overline{x}}{10}$, this gives

$$-100n\left(\frac{1}{\overline{x}} + \frac{1}{10-\overline{x}}\right) = -\frac{1000n}{(10-\overline{x})\overline{x}},$$

which is clearly < 0. So the MLE for p is $\widehat{p} = \frac{\overline{x}}{10} = 0.257$.

■

Estimation

Null Hypothesis (H_0) – Hypothesis that does not change current belief; Alternative Hypothesis (H_1) – Supports your claim.

10.1.1 Normal Distribution with Known Variance

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{N}(\mu,\sigma^2)$ with σ^2 known and μ unknown. We may wish to test if $\mu = \mu_0$ for some specific value μ_0 (e.g. $\mu_0 = 0$, $\mu_0 = 9.8$).

Then we can state our null and alternative hypotheses as

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \ne \mu_0.$$

Under $H_0 : \mu = \mu_0$, we then know both μ and σ^2. So for the sample mean $\overline{X}$ we have a known distribution for the test statistic

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim \Phi.$$

So if we define our rejection region R to be the $100\alpha\%$ tails of the standard normal distribution,

$$R = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right)$$
$$\equiv \left\{z \,\middle|\, |z| > z_{1-\frac{\alpha}{2}}\right\},$$

we have $\mathrm{P}(Z \in R|H_0) = \alpha$.

We thus reject H_0 at the $100\alpha\%$ significance level $\iff$ our observed test statistic

$$z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \in R.$$

10.1.2 Normal Distribution with Unknown Variance

Similarly, if σ^2 in the previous example were unknown, we still have that

$$T = \frac{\overline{X}-\mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}.$$

So for a test of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \ne \mu_0$ at the α level, the rejection region of our observed test statistic $t = \frac{\overline{x}-\mu_0}{s_{n-1}/\sqrt{n}}$ is

$$R = \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right) \cup \left(t_{n-1,1-\frac{\alpha}{2}}, \infty\right)$$
$$\equiv \left\{t \,\middle|\, |t| > t_{n-1,1-\frac{\alpha}{2}}\right\}.$$

Again, we have that $P(T \in R|H_0) = \alpha$.

10.2.2 Normal Distributions with Known Variances

Suppose

- $\underline{X} = (X_1, \ldots, X_{n_1})$ are i.i.d. $\mathrm{N}(\mu_X, \sigma_X^2)$ with μ_X unknown;
- $\underline{Y} = (Y_1, \ldots, Y_{n_2})$ are i.i.d. $\mathrm{N}(\mu_Y, \sigma_Y^2)$ with μ_Y unknown;
- the two samples $\underline{X}$ and $\underline{Y}$ are independent.

Then we still have that, independently,

$$\overline{X} \sim \mathrm{N}\left(\mu_X, \frac{\sigma_X^2}{n_1}\right), \qquad \overline{Y} \sim \mathrm{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right)$$

From this it follows that the difference in sample means,

$$\overline{X} - \overline{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right),$$

and hence

$$\frac{(\overline{X}-\overline{Y}) - (\mu_X-\mu_Y)}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

So under the null hypothesis $H_0 : \mu_X = \mu_Y$, we have

$$Z = \frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

So if σ_X^2 and σ_Y^2 are known, we immediately have a test statistic

$$z = \frac{\overline{x}-\overline{y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}}$$

which we can compare against the quantiles of a standard normal.

That is,

$$R = \left\{z \,\middle|\, |z| > z_{1-\frac{\alpha}{2}}\right\},$$

gives a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \ne \mu_Y$ at the $100\alpha\%$ level.

10.2.3 Normal Distributions with Unknown Variances

On the other hand, suppose σ_X^2 and σ_Y^2 are unknown. Then if we know $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ but σ^2 is unknown, we can still proceed.

We have

$$\frac{(\overline{X}-\overline{Y}) - (\mu_X-\mu_Y)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \Phi,$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \Phi.$$

but with σ unknown.

We need an estimator for the variance using samples from two populations with different means. Just combining the samples together into one big sample would over-estimate the variance, since some of the variability in the samples would be due to the difference in μ_X and μ_Y.

So we define the **bias-corrected pooled sample variance**

$$S_{n_1+n_2-2}^2 = \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2},$$

which is an unbiased estimator for σ^2.

We can immediately see that $s_{n_1+n_2-2}^2$ is indeed an unbiased estimate of σ^2 by noting

$$S_{n_1+n_2-2}^2 = \frac{n_1-1}{n_1+n_2-2}S_{n_1-1}^2 + \frac{n_2-1}{n_1+n_2-2}S_{n_2-1}^2;$$

That is, $s_{n_1+n_2-2}^2$ is a weighted average of the bias-corrected sample variances for the individual samples $\underline{x}$ and $\underline{y}$, which are both unbiased estimates for σ^2.

Then substituting $S_{n_1+n_2-2}$ in for σ we get

$$\frac{(\overline{X}-\overline{Y}) - (\mu_X-\mu_Y)}{S_{n_1+n_2-2}\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2},$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$T = \frac{\overline{X}-\overline{Y}}{S_{n_1+n_2-2}\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

So we have a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \ne \mu_Y$ at the $100\alpha\%$ level given by

$$R = \left\{t \,\middle|\, |t| > t_{n_1+n_2-2,1-\frac{\alpha}{2}}\right\},$$

for the statistic

$$t = \frac{\overline{x}-\overline{y}}{s_{n_1+n_2-2}\sqrt{1/n_1 + 1/n_2}}.$$

(1ai) Using Bayes' Theorem.

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

Calculate denominator:

$$P(x) = \int_0^1 \theta^{x}(1-\theta)^{n-x} \cdot \mathrm{Beta}(\theta; a, b)\,d\theta = \frac{1}{B(a,b)} \cdot \int_0^1 \theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta$$

Finding posterior distribution:

$$P(\theta|x) = \frac{\theta^{a+x-1}(1-\theta)^{b+n-x-1}}{\int_0^1 \theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta} = \mathrm{Beta}(\theta; a+x, b+n-x) /\!/$$

$$\hat{\theta}_B = \frac{x+a}{n+a+b} = \mu_{a+x,\, b+n-x} /\!/$$

$$\hat{\theta}_{MAP} = \frac{x+a-1}{n+a+b-2} = m_{a+x,\, b+n-x} /\!/$$

$$\hat{\theta}_{MLE} = \frac{x}{n} /\!/$$

P.T.O. →

Suppose we have a null hypothesis $H_0 : \theta = \theta_0$ for the value of the unknown parameter(s). Then under H_0 we know the pmf $\{p_j\}$, and so we are able to calculate the expected frequency counts $\underline{E} = (E_1, \ldots, E_k)$ by $E_j = np_j$. (Note again we have $\sum_{j=1}^k E_j = n$.)

We then seek to compare the observed frequencies with the expected frequencies to test for **goodness of fit**.

To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \ne \theta_0$ we use the **chi-square statistic**

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

If H_0 were true, then the statistic X^2 would approximately follow a **chi-square distribution** with $\nu = k - p - 1$ degrees of freedom.

- k is the number of values (categories) the simple random variable X can take.
- p is the number of parameters being estimated ($\dim(\theta)$).
- For the approximation to be valid, we should have $\forall j, E_j \ge 5$. This may require some merging of categories.

$$\hat{\theta}_B = \frac{18+15}{30+15+10} = 0.6 /\!/ \quad \hat{\theta}_{MAP} = \frac{\quad}{30+15+10-2} \approx 0.604 \text{ (3sf)} /\!/$$

$$\hat{\theta}_{maxPrior} = \frac{15-1}{15+10-2} = \frac{14}{23} \approx 0.609 \text{ (3s.f.)} /\!/$$

$$\hat{\theta}_{meanPrior} = \frac{15}{15+10} = 0.6 /\!/ \quad \hat{\theta}_{MLE} = \frac{18}{30} = 0.6 /\!/$$

So $\hat{\theta}_B = 0.6$, $\hat{\theta}_{MAP} = 0.604$, $\hat{\theta}_{MLE} = 0.6$, $\hat{\theta}_{maxPrior} = 0.609$, $\hat{\theta}_{meanPrior} = 0.6 /\!/$