Further Mathematics Statistics Template

Distributions

Combining Variables

For two separate variables X and Y, we could have Linear combination $Z = aX + bY$

Hence that we can obtain:

- $E(Z) = E(aX + bY) = aE(X) + bE(Y)$
- $Var(Z) = a^2Var(X) + b^2Var(Y)$

Therefore, if we combined two variable $X \sim \mathcal{N}(\bar{x}, \sigma_x^2)$ and $Y \sim \mathcal{N}(\bar{y}, \sigma_y^2)$ as $Z = aX + bY$, we would have that $Z \sim \mathcal{N}(a\bar{x} + b\bar{y}, a^2\sigma_x^2 + b^2\sigma_y^2)$

Similarly, if we combined two variable $X \sim Po(\lambda_x)$ and $Y \sim Po(\lambda_y)$ as $Z = aX + bY$, we would have that $Z \sim Po(a\lambda_x + b\lambda_y)$

Notice that aX is not the same as $\underbrace{X + X + \cdots + X}_{a}$, the lateral one means all of the variables are

independent of each other.

PDF and CDF

Relationship between PDF and CDF

a PDF describing the probability for a distribution X can be expressed as:

$$f(x) = \begin{cases} p(x) & a < x < b \\ q(x) & b < x < c \\ \cdots & \\ 0 & otherwise \end{cases}$$

$P(a < x < b) = \int_a^b f(x)dx$, note that individual $P(X = a)$ for PDF is always 0

Note that $\int_{-\infty}^{\infty} f(x)dx = 1$ and $f(x) \geq 0$

so that the CDF for the given PDF can be expressed as:

$$F(x) = \begin{cases} 0 & x < a \\ P(x) = \int p(x)dx & a < x \leq b \\ Q(x) = \int q(x)dx & b < x \leq c \\ \cdots & \\ 1 & c < x \end{cases}$$

$P(x < a) = F(a)$

Note that $F(x)$ is increase and that it reached 1 after threshold value

($p(x), q(x), P(x), Q(x)$ are expression of x)

Calculating Statistics

- $E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
- $Var[g(x)] = \int_{-\infty}^{\infty} g(x)^2 f(x)dx - E(g(x))^2$
- Median x is that $\int_{-\infty}^{x} f(u)du = \int_{x}^{\infty} fu)du = 0.5$. also $F(x) = 0.5$
- Mode x is the $argmax_x f(x)$, which means that x has highest magnitude on $f(x)$

Changing Variable

For a substitution of variable $Y = p(X)$ for the CDF $F(X)$, we wished to obtain the correspondence CDF $G(Y)$, we want to consider the approach by considering the definition of CDF

First we want to Obtain $X = q(Y)$ given that $Y = p(X)$

Note that $p[.]$ and $q[.]$all represent the relationship between two variables

$G(Y) = P(Y < y) = P(p(X) < y) = P(x < q(y)) = F(q(y))$

Notice during this process some expression such as $-x < y$ the expression should be changed to $1 - P(x < X)$ and hence the CDF should also be changed subsequently

hence you would be able to evaluate $G(y)$

the PDF $g(y)$ could be obtain by taking $\frac{d}{dx}G(y)$

Poisson distribution

$X \sim Po(\lambda)$ where λ is the mean number of event in given interval

this is used to find

Condition for Poisson distribution

- event occur randomly in space or time
- event occur singly(cannot be simultaneously)
- event occur independently
- event occur at a constant rate

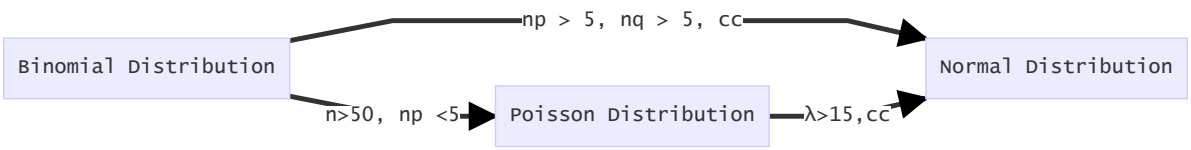
Probability

- $P(X = n) = \frac{e^{-\lambda}\lambda^n}{n!}$ for $n \in \mathbb{N}$
- $P(a \leq X \leq b) = \sum_{x=a}^{b} \frac{e^{-\lambda}\lambda^x}{x!}$
- $P(X > n) = 1 - P(x \leq n) = 1 - \sum_{x=0}^{n} \frac{e^{-\lambda}\lambda^n}{n!}$

Means and Variance

- $E(x) = \lambda$
- $Var(x) = \lambda$

Approximate Poisson distribution with other distribution



Continuity Correction

when changing variable from discreate to continuous, continuity correction is required to ensure the range cover all interval:

Discreate	Continuous
$X = a$	$a - 0.5 \leq X \leq a + 0.5$
$X > a$	$X \geq a + 0.5$
$X \geq a$	$X \geq a - 0.5$
$X < a$	$X \leq a - 0.5$
$X \leq a$	$X \leq a + 0.5$

Binomial to Poisson

if $n > 50$, $np < 5$ we could approximate binomial distribution using Poisson distribution

Notice you should verify the condition before you conduct the approximation

Specifically, we can have $X \sim B(n, p) \rightarrow Y \sim Po(np)$

Poisson to Normal

if $\lambda > 15$, we could approximate Poisson distribution using Normal distribution

Notice you should verify the condition before you conduct the approximation

Specifically, we can have $X \sim Po(\lambda) \rightarrow Y \sim \mathcal{N}(\lambda, \lambda)$

When obtaining the probability, you should use the continuity correction

Binomial to Normal

If $np > 5$ and $nq > 5$ we could approximate binomial distribution using Normal distribution

Notice you should verify the condition before you conduct the approximation

Specifically, we can have $X \sim B(n, p) \rightarrow Y \sim \mathcal{N}(np, npq)$

When obtaining the probability, you should use the continuity correction

Geometric distribution

$X \sim Geo(p)$ where p is the probability of success for one trail

this is used to find the number of attempt that leads to the first success

Condition

- Only two possible outcomes(success/failure)
- Probability of success p is constant
- Each event is independent

Probability

- $P(X = n) = (1-p)^{n-1}p$
- $\begin{cases} P(X \leq x) = 1-(1-p)^x \\ P(X > x) = (1-p)^x \end{cases}$
- $\begin{cases} P(X < x) = 1-(1-p)^{x-1} \\ P(X \geq x) = (1-p)^{x-1} \end{cases}$

Means and Variance

- $E(x) = \dfrac{1}{p}$
- $Var(x) = \dfrac{q}{p^2}$

Negative exponential distribution

$X \sim Exp(\lambda)$ where λ is the average

this is used to model the duration of the event

Probability

- $P(a < X < b) = e^{-\lambda a} - e^{-\lambda b}$
- $P(X < x) = 1 - e^{-\lambda x}$
- $P(X > x) = e^{-\lambda x}$

Means and Variance

- $E(x) = \dfrac{1}{\lambda}$
- $Var(x) = \dfrac{1}{\lambda^2}$

Memoryless

This probability distribution have a very special property, which is **memoryless**

This property suggest that the duration of the current event is irrelevant to the previous event:

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

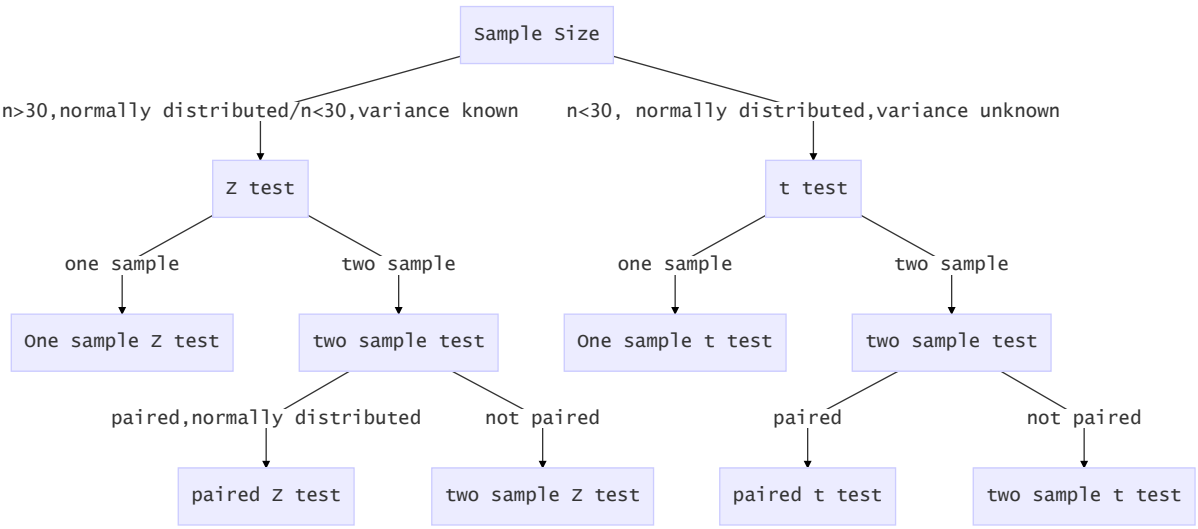
Please be aware of this property in the context of the question

PDF and CDF

- $f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$ (Notice you can derive this from CDF)
- $F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$ (Notice you should be able to obtain this through cumulative probability)

Inference with Normal and t distributions

The decision Process on Choosing the type of Hypothesis test



if the variance σ^2 is given, then it can used in the z test. Otherwise, the unbiased estimator $\hat{\sigma}^2$ should be calculated. All numerical value involved in the test shall be calculated before conducting the test:

- sample mean: $\bar{x} = \dfrac{\sum x}{n_x}$
- unbiased sample variance: $\hat{\sigma_x}^2 = \dfrac{n}{n-1}\left\{\dfrac{\sum x^2}{n} - \bar{x}^2\right\}$ (this is usually preferred in the FMA context for hypothesis test)

One Sample Test

Define the variable[if needed]

Hypothesis

H_0: $\mu_x = a$

H_1: $\mu_x \neq a, \mu_x > a, \mu_x < a$

Test Statistics

Note that you should state CLT is used if only $n > 30$ is given

$$Z = \frac{\bar{x} - \mu_x}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x}}} \text{ or } T = \frac{\bar{x} - \mu_x}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x}}}$$

if it is a two tailed test, then obtain $z_{\alpha/2}$ or $t_{\alpha/2,n-1}$, else obtain z_α or $t_{\alpha,n-1}$(Note DGF = $n-1$)

if $|Z| < z$ or $|T| < t$, H_0 accepted, otherwise H_0 rejected

Conclusion

Given the current data, there {is/is not} evidence to say that at α% significant level, {description of μ_x hypothesis}

Confidence Interval

For Z test that the variance is known: $[\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n}}]$

For Z test that the variance is unknown: $[\bar{x} - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_x^2}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_x^2}{n}}]$

For t test: $[\bar{x} - t_{\alpha/2,n-1}\sqrt{\frac{\hat{\sigma}_x^2}{n}}, \bar{x} + t_{\alpha/2,n-1}\sqrt{\frac{\hat{\sigma}_x^2}{n}}]$

Two sample test

Define the variable[if needed]

Hypothesis

H_0: $\mu_x - \mu_y = a$

H_1: $\mu_x - \mu_y \neq a / \mu_x - \mu_y > a / \mu_x - \mu_y < a$

Test Statistics

pooled estimate is required if n_1, n_2 < 30, but it is always recommended unless question specified not to use

$$\hat{\sigma^2} = \frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}$$

Note this formula used unbiased estimator of variance, the MF10 one is the one with biased estimator

For known variance Z test: $Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$

For unknown variance Z or T test: $Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}^2}{n_x} + \frac{\hat{\sigma}^2}{n_y}}}$ or $T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}^2}{n_x} + \frac{\hat{\sigma}^2}{n_y}}}$

if it is a two tailed test, then obtain $z_{\alpha/2}$ or $t_{\alpha/2,n-2}$, else obtain z_α or $t_{\alpha,n-2}$(Note DGF = $n-2$)

if $|Z| < z$ or $|T| < t$, H_0 accepted, otherwise H_0 rejected

Conclusion

Given the current data, there {is/is not} evidence to say that at α% significant level, {description of μ_x, u_y hypothesis}

Confidence Interval

For Z test with known variance: $[(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}, (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}]$

For Z test with pooled variance: $[(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n_x}+\frac{\hat{\sigma}^2}{n_y}}, (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n_x}+\frac{\hat{\sigma}^2}{n_y}}]$

For t test: $[(\bar{x}-\bar{y}) - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{n_x}+\frac{\hat{\sigma}^2}{n_y}}, (\bar{x}-\bar{y}) + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{n_x}+\frac{\hat{\sigma}^2}{n_y}}]$

Paired Test

for a paired data X and Y, you may conduct a test for the difference D between X and Y:

X	x_1	x_2	$\dots$	x_i
Y	y_1	y_2	$\dots$	y_i
$D = X - Y$	$x_1 - y_1$	$x_2 - y_2$	$\dots$	$x_i - y_i$

Define the variable[if needed]

Hypothesis

H_0: $\mu_D =$a

H_1: $\mu_D \neq a, \mu_D > a, \mu_D < a$

Test Statistics

$Z = \frac{\bar{D}-\mu_D}{\sqrt{\frac{\hat{\sigma}_D^2}{n_D}}}$ or $T = \frac{\bar{D}-\mu_D}{\sqrt{\frac{\hat{\sigma}_D^2}{n_D}}}$

if it is a two tailed test, then obtain $z_{\alpha/2}$ or $t_{\alpha/2,n-1}$, else obtain z_α or $t_{\alpha,n-1}$(Note DGF = $n-1$)

if $|Z| < z$ or $|T| < t$, H_0 accepted, otherwise H_0 rejected

Conclusion

Given the current data, there {is/is not} evidence to say that at α% significant level, {description of μ_D hypothesis}

Confidence Interval

For Z test: $[\bar{D} - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_D^2}{n}}, \bar{D} + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_D^2}{n}}]$

For t test: $[\bar{D} - t_{\alpha/2,n-1}\sqrt{\frac{\hat{\sigma}_D^2}{n}}, \bar{D} + t_{\alpha/2,n-1}\sqrt{\frac{\hat{\sigma}_D^2}{n}}]$

Chi Square Test

Goodness of fit test

Hypothesis

H_0: {distribution} is a good model to fit the data

H_1: {distribution} is not a good model to fit the data

Frequency Table

Number	Interval
Observed	...
Expected	...
$\frac{(O-E)^2}{E}$	...

Calculating Expected Value

Expected Value can be calculated given the following situation:

- PDF $f(x)$ given, $N = N_{total} f(n)$ for each interval
- Poisson distribution given, the first $n-1$ term should be obtain from $\frac{\lambda^x e^{-\lambda}}{x!}$, and the n^{th} term should be $1 - P(x < n)$
- Binomial distribution given, the term should be calculated as $nCr * p^r (1-p)^{(n-r)}$
- Normal distribution given, the term should be $N * P(n_1 < x < n_2)$

Note that before calculating the χ score the expected value should be combined to the interval next to it so that each interval for expected value >5

Calculating the Test Statistics

$\chi^2 = \sum \frac{(O-E)^2}{E}$

$DOF = N_{given\ outcome} - N_{parameters\ calculated} - 1$, for example:

- for normal distribution if $\bar{x}$ and $\hat{\sigma_x}^2$ predicted from data, $N_{parameters\ calculated} = 2$, and so on
- for Poisson distribution if λ predicted from data, $N_{parameters\ calculated} = 1$
- for Binomial distribution if p predicted from data, $N_{parameters\ calculated} = 1$

Obtain the $\mathrm{X}^2_{DOF,\alpha}$ value from the MF10 booklet, if $\chi^2 > \mathrm{X}^2_{DOF,\alpha}$, then H_0 rejected. Otherwise H_0 accepted

Conclusion

Given the current evidence, there is {enough/not enough} evidence to say that at α significant level the model{distribution} is a good model for the data

Test for Independent

Hypothesis

H_0: {variable I} is independent of {Variable II}

H_1: {variable I} is not independent of {Variable II}

Contingency Table

For observed frequency, we would have the table here which the sum of frequency we could obtained from the observed value:

Observed Frequency	x_1	x_2	$\dots$	x_i	**Total**
y_1					$\sum f_{i,1}$
y_2					$\sum f_{i,1}$
$\dots$					$\dots$
y_j					$\sum f_{i,j}$
Total	$\sum f_{1,j}$	$\sum f_{2,j}$	$\dots$	$\sum f_{i,j}$	$\sum f$

Hence we would be able to calculate the expected frequency from the contingency table:

Expected Frequency	x_1	x_2	$\dots$	x_i	**Total**
y_1	$f_{total} * \frac{f_{1,j}}{f_{total}} * \frac{f_{i,1}}{f_{total}}$	$f_{total} * \frac{f_{2,j}}{f_{total}} * \frac{f_{i,1}}{f_{total}}$	$\dots$	$f_{total} * \frac{f_{i,j}}{f_{total}} * \frac{f_{i,1}}{f_{total}}$	$\sum f_{i,1}$
y_2	$f_{total} * \frac{f_{1,j}}{f_{total}} * \frac{f_{i,2}}{f_{total}}$	$f_{total} * \frac{f_{2,j}}{f_{total}} * \frac{f_{i,2}}{f_{total}}$	$\dots$	$f_{total} * \frac{f_{i,j}}{f_{total}} * \frac{f_{i,2}}{f_{total}}$	$\sum f_{i,1}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_j	$f_{total} * \frac{f_{1,j}}{f_{total}} * \frac{f_{i,j}}{f_{total}}$	$f_{total} * \frac{f_{2,j}}{f_{total}} * \frac{f_{i,j}}{f_{total}}$	$\dots$	$f_{total} * \frac{f_{i,j}}{f_{total}} * \frac{f_{i,j}}{f_{total}}$	$\sum f_{i,j}$
Total	$\sum f_{1,j}$	$\sum f_{2,j}$	$\dots$	$\sum f_{i,j}$	$\sum f$

Test Statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$DOF = (row - 1)(col - 1)$$

Obtain the $\mathrm{X}^2_{DOF,\alpha}$ value from the MF10 booklet, if $\chi^2 > \mathrm{X}^2_{DGF,\alpha}$, then H_0 rejected. Otherwise H_0 accepted

Conclusion

Given the current evidence, there is {enough/not enough} evidence to say that at α significant level the two variables are independent

Note that if H_0 rejected, we could only say that the two variable is **associated** but we cannot say they are correlated

Bivariate Data

Given two variable y and x, if we attempt to find a correlation between them, it is called regression. In FMA we are attempting to find the best linear regression line between the two variable. This can be done via the measure of least square, which is the sum of square of vertical distance from the point to the regression line. This method is also known as OLS(ordinary least square).

[Side note][https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem]: The reason we prefer OLS is that it is the best unbiased estimator with least variance for the linear model

Obtaining the regression line

Given $\hat{y} = a + bx$ as the regression line, we attempt to find (a, b) such that $argmin_{(a,b)} \sum [(bx_i + a) - y_i]^2$
This could be done by letting $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$

Hence we obtained the expression for the coefficient for the regression line:

$$b = \frac{S_x y}{S_x x} = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\frac{\sum x^2}{n} - \bar{x}^2} \text{ and } a = \bar{y} - b\bar{x}$$

Notice that the regression line always past $\bar{x}$ and $\bar{y}$

This is the coefficient for the regression line for y on x(taking x as independent and y as dependent). Note during the exam you should plug in the real value to the examiner to demonstrate you actually calculate this

Similarly we can have regression line for x on y (taking y as independent and x as dependent), which is $\hat{x} = c + dy$

the coefficient could be obtained similarly by using OLS:

$$d = \frac{S_x y}{S_x x} = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\frac{\sum x^2}{n} - \bar{x}^2} \text{ and } c = \bar{x} - d\bar{y}$$

Product moment correlation coefficient

The line of best fit can always be calculated, however the regression line may or may not reflect the real linear relationship within the data. Therefore we defined the measurement of PMCC to describe the nature of the linear correlation.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2 - \bar{x}^2}{n}}\sqrt{\frac{\sum y^2 - \bar{y}^2}{n}}}$$

which means that $r^2 = bd$ as b and d are coefficient for regression lines for y on x and for x on y

- b>0, $r = \sqrt{bd}$, the correlation is positive
- b<0, $r = -\sqrt{bd}$, the correlation is negative

Comment on PMCC

$r \in [-1, 1]$ and that:

Either

if $0.75 < |r| < 1$ then there is strong correlation

if $|r| < 0.75$ then there is not strong correlation

Or:

if $r > \rho_{\alpha\%,n}$ it is reliable

if $r > \rho_{\alpha\%,n}$, it is not reliable

Conducting Hypothesis Test on the PMCC

Hypothesis

non-zero correlation	positive correlation	negative correlation
H_0: $\rho = 0$, H_1: $\rho \neq 0$	H_0: $\rho = 0$, H_1: $\rho > 0$	H_0: $\rho = 0$, H_1: $\rho < 0$

The test for non-zero correlation is two tailed test, and the test for positive and negative correlation is one-tailed test.

Test statistics

The critical value for $\rho_{n,sig\ level}$ can be obtained from MF10, (notice the difference between two tailed and one tailed significant level). If $\rho < |r|$1 reject H_0, otherwise accept H_0,

Conclusion

Given the current evidence, there is {enough/not enough} evidence to say that at α significant level there is a {non-zero/positive/negative} correlation between the two variable