Formulas statistics

z-score: $z_i = \frac{x_i - \bar{x}}{s}$

$r = \frac{\sum^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum^n (x_i-\bar{x})^2 \sum^n (y_i-\bar{y})^2}}$ correlation $[-1, +1]$ $r^2 = \frac{Var(\hat{y})}{Var(y)}$

$\hat{y} = a + bx$ $\hat{y}$: predicted value $b = r\frac{s_y}{s_x} = \frac{\sum^n (x_i-\bar{x})(y_i-\bar{y})}{\sum^n (x_i-\bar{x})^2}$ $a = \bar{y} - b\bar{x}$

residual sum of squares: $\sum^n (y_i - \hat{y}_i)^2$

sampling distribution of population proportion: $N \sim \left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)$ $\sim N(\mu, \sigma^2)$ by CLT for categorical if $n\hat{p}$ and $n(1-\hat{p}) \geq 15$

sampling distribution of sample mean: $N \sim \left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$ σ: population std, s: sampling std $\sim N(\mu, \sigma^2)$ by CLT when $n > 30$ for quantitative

CI of population proportion (estimate): $\hat{p} \pm z_{\alpha/2}(se)$; $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ($1-\alpha$ confidence level)

CI of population mean (estimate): $\bar{x} \pm z_{\alpha/2}(se)$; $se = \frac{s}{\sqrt{n}}$ with n large or underlying normally distributed population

$\bar{x} \pm t_{df, \alpha/2}(se)$; $se = \frac{s}{\sqrt{n}}$ $df = n-p$ with small n; normallity assumption with n large: CLT

CI for two means:
two sample pooled t-interval $(\bar{x}-\bar{y}) \pm t_{\alpha/2, n+m-2}\, S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$; S_p^2 is the pooled sample variance

$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{m+n-2}$ (unbiased estimator of common σ^2)

two sample t-interval (different σ^2) $(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}$

CI for difference in two population proportions $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Significance test
1. assumptions
2. hypotesis
3. test statistic
4. p-value
5. conclusion

↓ reject H_0 if p-value < α

proportions
categorical 0-1, sample size large ~ normal
$H_0: p = p_0$; $H_1: p < p_0$ / $p \neq p_0$ / $p > p_0$
$z_{obs} = (\hat{p} - p_0) / \sqrt{\frac{p_0(1-p_0)}{n}}$ $\left[\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\right]$
$H_1: p < p_0 \Rightarrow P(z < z_{obs})$
$H_1: p \neq p_0 \Rightarrow P(|t| > |z_{obs}|) = 2 \cdot P(z > |z_{obs}|)$
$H_1: p > p_0 \Rightarrow P(z > z_{obs})$

means
quantitative, population ~ normal
$H_0: \mu = \mu_0$; $H_1: \mu < \mu_0$ / $\mu \neq \mu_0$ / $\mu > \mu_0$
$T_{obs} = (\bar{x} - \mu_0) / \sqrt{s^2/n}$ $\left[\frac{\bar{x}-\mu_0}{s}\sqrt{n}\right]$
$H_1: \mu_0 < \mu_0 \Rightarrow P(z < T_{obs})$
$H_1: \mu_0 \neq \mu_0 \Rightarrow P(|t| > |T_{obs}|) = 2 \cdot P(z > |T_{obs}|)$
$H_1: \mu > \mu_0 \Rightarrow P(z > T_{obs})$

$\bar{y}$ mean of a sample
μ; $\mu = E(y)$ mean of a population $E(y) = \beta_0 + \beta_1 x$ b_0; b_1 in the sample
$y = \beta_0 + \beta_1 x + \varepsilon$; $\varepsilon \sim N(0, \sigma^2)$ is the simple linear regression model
fitted model: $\hat{y}_i = b_0 + b_1 x_i$

$SSE = \sum^n (y_i - \hat{y}_i)^2$ sum of squared errors (residuals)

$b_1 = \frac{\sum^n (x_i-\bar{x})(y_i-\bar{y}_i)}{\sum^n (x_i-\bar{x})^2}$ $b_0 = \bar{y} - b_1\bar{x}$ estimates of β_1 and β_0

$b_1 \sim N(\beta_1, Var(b_1))$ $Var(b_1) = \frac{\sigma^2}{\sum^n (x_i-\bar{x})^2}$
$b_0 \sim N(\beta_0, Var(b_0))$ $Var(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum^n (x_i-\bar{x})^2}\right]$

$SSE = \sum^n (y_i - \hat{y}_i)^2$

$MSE = \frac{SSE}{n-p}$

$\frac{b_1 - \beta_1}{\sqrt{Var(b_1)}} \sim N(0,1)$ $\frac{b_1 - \beta_1}{S(b_1)} \sim t(n-p)$

$S = \sqrt{MSE}$

$\frac{b_0 - \beta_0}{\sqrt{Var(b_0)}} \sim N(0,1)$ $\frac{b_0 - \beta_0}{S(b_0)} \sim t(n-p)$

$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$ $r = \sqrt{R^2}$

$SST = \sum^n (y_i - \bar{y})^2$

$SST = SSR + SSE \rightarrow SSR = SST - SSE$

$SSR = \sum^n (\hat{y}_i - \bar{y})^2$ $MSR = \frac{SSR}{p-1}$

Hypothesis testing

1. state null and alternative Hypothesis
2. test statistic
3. conclusion : critical value approach
   p-value

two-sided test

$H_0: \beta = 0$; $H_1: \beta \neq 0$

$T = \frac{b}{S(b)} \sim t(n-p)$ $\left[\frac{b-\beta}{S(b)} \sim t(n-p) \right]$

note: $p = 2$ (only β_0 and β_1) simple linear regression

crit. val: if $|T_0| > t_{\alpha/2}(n-p)$ reject H_0

p-val: if p-value $< \alpha$ reject H_0

p-val $= P(|T| > |T_0|)$

(one sided test) : reject H_0 when $T_0 \gtrless t_{\alpha/2}(n-p)$ or $P(T \gtrless T_0) < \alpha$

$H_1: \beta \gtrless 0$

$(t\text{-value})^2 = (F\text{-value})$ simple linear regression

$F(1, n-1)$

Confidence interval for slope β_1

$b_1 \pm t_{\alpha/2}(n-2) \cdot s(b_1)$

$\hookrightarrow \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$

CI for $E(Y)$:

$\hat{Y} \pm t_{(1-\alpha)/2}(n-2) \cdot s\{\hat{Y}\}$

$\hat{Y} = b_0 + b_1 x$

$s\{\hat{Y}\} = \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$

PI for y: (prediction interval)

$\hat{y} \pm t_{\alpha/2}(n-2) \sqrt{s^2\{\hat{y}\} + MSE}$

Multiple linear regression

$E(Y) = X\beta$ regression equation $\quad Y = E(Y) + \underset{\text{residual}}{\varepsilon}$ regression model

estimate of vector β : $b = (X^TX)^{-1}X^TY \sim MVN(\beta, \sigma^2(X^TX)^{-1})$

$\hat{Y} = Xb$ fitted model

$SSE = e^Te = [(Y-Xb)^T(Y-Xb)] = \sum (Y_i - \hat{Y}_i)^2$

$MSE = \frac{SSE}{n-p}$

$S = \sqrt{MSE}$

$SST = \sum (Y_i - \bar{Y})^2$

$SSR = SST - SSE$

$MSR = \frac{SSR}{p-1}$ $\quad R^2 = \frac{SSR}{SSE}$

in hypothesis testing : $T = \frac{\text{sample coefficient}}{\text{std error of the coefficient}} \sim t(n-p)$ under H_0

ANOVA table (decomposition of variance)

source	DF	SS	MS	F
regression	$p-1$	SST-SSE	MSR	$\frac{MSR}{MSE}$
error	$n-p$	SSE	MSE	
total	$n-1$	SST		

(significance test)

F-test $H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

H_1: at least one $\beta_i \neq 0$ for $i = 1, \ldots, p-1$

$F = \frac{MSR}{MSE} \sim F(p-1, n-p)$ if $F_0 > F_\alpha(p-1, n-p)$ reject H_0

if $P(F > F_0) < \alpha$ reject H_0

t-test tests linear relation between Y and a certain X_i while all other x-variables are in the model

F-test tests linear relation between Y and all x-variables together