Properties of expectation, covariance and variance

$$\begin{aligned}
\mathbb{E}(X+Y) &= \mathbb{E}(X)+\mathbb{E}(Y)\\
\mathbb{E}(aX+b) &= a\mathbb{E}(X)+b\\
\mathbb{C}(X+Y,Z) &= \mathbb{C}(X,Z)+\mathbb{C}(Y,Z)\\
\mathbb{C}(aX+b,cY) &= ac\mathbb{C}(X,Y)\\
\mathbb{V}(aX+bY+c) &= a^2\mathbb{V}(X)+b^2\mathbb{V}(Y)+2ab\mathbb{C}(X,Y)
\end{aligned}$$

If $X\sim\mathcal{N}\left(\mu_x,\sigma_x^2\right)$ and $Y\sim\mathcal{N}\left(\mu_y,\sigma_y^2\right)$ and $\sigma_{x,y}=\mathbb{C}(X,Y)$

$$aX+bY\sim\mathcal{N}\left(a\mu_x+b\mu_y\,,\,a^2\sigma_x^2+b^2\sigma_y^2+2ab\sigma_{x,y}\right)$$

Central limit theorem. If $X_1,\dots,X_n$ are iid random variables with expected value μ and variance σ^2, then

$$Z_n=\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\xrightarrow{d}Z\sim\mathcal{N}(0,1)\qquad \bar{X}_n\approx\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$$

Let Z and Y be two independent r.v. $Z\sim\mathcal{N}(0,1)$ and $Y\sim\chi^2_n$ then $T_n=\frac{Z}{\sqrt{Y/n}}\sim t_n$

Let X and Y be two independent r.v. with $X\sim\chi^2_{n_1}$ and $Y\sim\chi^2_{n_2}$. Then $F=\frac{X/n_1}{Y/n_2}\sim F_{n_1,n_2}$.

The gamma function is $\Gamma(z):=\int_0^{+\infty}t^{z-1}e^{-t}dt$, satisfies $\Gamma(z+1)=z\Gamma(z)$, and, $\forall n\in\mathbb{N}$, we have $\Gamma(n+1)=n!$

The moment generating function of a r.v. X is defined as $m_X(t):=E[e^{tX}]$. It has the following properties

$$m_{aX+b}(t)=e^{bt}m_X(at)$$

$$m_{S_n}(t)=\prod_{i=1}^n m_{X_i}(t)$$

where $S_n=\sum_{i=1}^n X_i$ and the X_i's are independent r.v. The moment generating function characterizes the distribution and helps us retrieve the moments, i.e. $E[X^k]=\left.\frac{\partial^k m_X(t)}{\partial t^k}\right|_{t=0}$

Properties of $\bar{X}$ and S^2, 1 population

	Distribution of X	μ	σ^2	Distribution result
$\bar{X}$	unknown and large n	ok	ok	$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\sim\mathcal{N}(0,1)$
S^2	$\mathcal{N}\left(\mu,\sigma^2\right)$	-	ok	$(n-1)\frac{S^2}{\sigma^2}\sim\chi^2_{n-1}$
$\bar{X}$	$\mathcal{N}\left(\mu,\sigma^2\right)$	ok	-	$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}}\sim t_{n-1}$

Properties of $\bar{X}$ and S^2, 2 normal populations

$$S^2_{pool}=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\,,$$

		μ_1,μ_2	σ_1^2,σ_2^2	Statistics
$\bar{X}_1-\bar{X}_2$	$\sigma_1^2\neq\sigma_2^2$	ok	ok	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim\mathcal{N}(0,1)$
S^2_{pool}	$\sigma_1^2=\sigma_2^2$	-	ok	$(n_1+n_2-2)\frac{S^2_{pool}}{\sigma^2}\sim\chi^2_{n_1+n_2-2}$.
S_1^2,S_2^2	$\sigma_1^2\neq\sigma_2^2$	-	ok	$\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}\sim F_{n_1-1\,,\,n_2-1}$
$\bar{X}_1-\bar{X}_2$	$\sigma_1^2=\sigma_2^2$	ok	-	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_{pool}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2}$

In what follows, the notation $t_{n-1\;\alpha}$ denotes the α-quantile of a Student distribution with $n-1$ degrees of freedom. The same notation is used for the Chi-squared and Fisher distribution.

Single mean test, $H_0:\mu=\mu_0$. We reject H_0 at the level α if $T(\boldsymbol{x})=\sqrt{n}\frac{\bar{x}-\mu_0}{s}$

a) $H_1:\ \mu>\mu_0\qquad T(\boldsymbol{x})>t_{n-1\;1-\alpha}$

b) $H_1:\ \mu<\mu_0\qquad T(\boldsymbol{x})<t_{n-1\;\alpha}$

c) $H_1:\ \mu\neq\mu_0\qquad T(\boldsymbol{x})<t_{n-1\;\alpha/2}$ or $T(\boldsymbol{x})>t_{n-1\;1-\alpha/2}$

Single variance test, $H_0:\sigma^2=\sigma_0^2$. We reject H_0 at the level α if $T(\boldsymbol{x})=(n-1)\frac{s^2}{\sigma_0^2}$

a) $H_1:\sigma^2\neq\sigma_0^2\qquad T(\boldsymbol{x})<\chi^2_{n-1\;\alpha/2}$ or $T(\boldsymbol{x})>\chi^2_{n-1\;1-\alpha/2}$

b) $H_1:\sigma^2>\sigma_0^2\qquad T(\boldsymbol{x})>\chi^2_{n-1\;1-\alpha}$

c) $H_1:\sigma^2<\sigma_0^2\qquad T(\boldsymbol{x})<\chi^2_{n-1\;\alpha}$

Test on 2 sample means. $T(\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{(\bar{X}_1-\bar{X}_2)-\delta}{S_{pool}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$. We reject $H_0:\mu_1-\mu_2=\delta$ at the level α if

$$
\begin{array}{llll}
a) & H_1: & \mu_1-\mu_2>\delta & T(\boldsymbol{x}_1,\boldsymbol{x}_2)>t_{n_1+n_2-2\,1-\alpha} \\
b) & H_1: & \mu_1-\mu_2<\delta & T(\boldsymbol{x}_1,\boldsymbol{x}_2)<t_{n_1+n_2-2\,\alpha} \\
c) & H_1: & \mu_1-\mu_2\neq\delta & \begin{cases} T(\boldsymbol{x}_1,\boldsymbol{x}_2)<t_{n_1+n_2-2\,\alpha/2} & or \\ T(\boldsymbol{x}_1,\boldsymbol{x}_2)>t_{n_1+n_2-2\,1-\alpha/2} \end{cases}
\end{array}
$$

Test on 2 variances. $T(\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{S_1^2}{S_2^2}\sim F_{n_1-1\,,\,n_2-1}$. We reject $H_0:\sigma_1^2=\sigma_2^2$ at the level α if

$$
\begin{array}{llll}
a) & H_1: & \sigma_1\neq\sigma_2 & \begin{cases} T(\boldsymbol{x}_1,\boldsymbol{x}_2)<F_{n_1-1,n_2-,1\,\alpha/2} & or \\ T(\boldsymbol{x}_1,\boldsymbol{x}_2)>F_{n_1-1,n_2-,1\,1-\alpha/2} \end{cases} \\
b) & H_1: & \sigma_1>\sigma_2 & T(\boldsymbol{x}_1,\boldsymbol{x}_2)>F_{n_1-1,n_2-,1\,1-\alpha} \\
c) & H_1: & \sigma_1<\sigma_2 & T(\boldsymbol{x}_1,\boldsymbol{x}_2)<F_{n_1-1,n_2-,1\,\alpha}
\end{array}
$$

Simple linear regression $Y_i=\beta_0+\beta_1x_i+\epsilon_i$, where $\epsilon_i\sim N(0,\sigma^2)$:

$$\widehat{\beta}_1=\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2}=\frac{S_{xy}}{S_{xx}}\qquad,\qquad\widehat{\beta}_0=\bar{y}-\widehat{\beta}_1\bar{x}$$

If $\overline{x^2}=\frac{1}{n}\sum_{i=1}^n x_i^2$, the variances of estimators are

$$\mathbb{V}\left(\widehat{\beta}_0\right)=\frac{\sigma^2\,\overline{x^2}}{S_{xx}}\qquad\mathbb{V}\left(\widehat{\beta}_1\right)=\frac{\sigma^2}{S_{xx}}\quad\mathbb{C}\left(\widehat{\beta}_0,\widehat{\beta}_1\right)=\frac{-\sigma^2\bar{x}}{S_{xx}}$$

Multiple linear regression $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is a $n\times(k+1)$ matrix and $\boldsymbol{\epsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$.

$$\widehat{\boldsymbol{\beta}}\;=\;\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\qquad\widehat{\boldsymbol{\beta}}\sim\mathcal{N}\left(\boldsymbol{\beta}\,,\sigma^2\,\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right)$$

An unbiased estimator of σ^2 is $\widehat{\sigma}^2=\frac{1}{n-(k+1)}\underbrace{\left(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^\top\left(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}_{SSE}$ and $(n-(k+1))\,\frac{\widehat{\sigma}^2}{\sigma^2}=\frac{SSE}{\sigma^2}\sim\chi^2_{n-(k+1)}$

$SST=\sum_{i=1}^n(y_i-\bar{y})^2=\boldsymbol{y}^\top\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{J}_n\right)\boldsymbol{y}$	$\frac{SST}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^n\left(Y_i-\bar{Y}\right)^2\sim\chi^2_{n-1}$
$SSR=\sum_{i=1}^n(\widehat{y}_i-\bar{y})^2=\boldsymbol{y}^\top\left(\boldsymbol{H}-\frac{1}{n}\boldsymbol{J}_n\right)\boldsymbol{y}$	$\frac{SSE}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^n\left(Y_i-\widehat{Y}_i\right)^2\sim\chi^2_{n-(k+1)}$
$SSE=\sum_{i=1}^n(y_i-\widehat{y}_i)^2=\boldsymbol{y}^\top\left(\boldsymbol{I}_n-\boldsymbol{H}\right)\boldsymbol{y}$	$\frac{SSR}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^n\left(\widehat{Y}_i-\bar{Y}\right)^2\sim\chi^2_k$

$$F^*=\frac{SSR\,/\,k}{SSE\,/\,(n-k-1)}\sim F_{k\,,\,n-(k+1)}$$

Test on significance of the model. We reject $H_0:\beta_1=\ldots=\beta_k=0$ against $H_1{:}\beta_j{\neq}0$ for some $j\in\{1,\ldots,k\}$, if

$$F^*>F_{k,n-(k+1)\;1-\alpha}$$

Test on regression coefficients. Let $c_{j,j}$ be the j^{th}diagonal element of $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$.

We reject $H_0:\beta_j{=}\beta_{j,0}$ at the level α if $T_j^*=\frac{\widehat{\beta}_j-\beta_{j,0}}{\widehat{\sigma}\sqrt{c_{jj}}}$

$$
\begin{array}{llll}
a) & H_1: & \beta_j{>}\beta_{j,0} & T_j^*>t_{n-k-1\;1-\alpha} \\
b) & H_1: & \beta_j{<}\beta_{j,0} & T_j^*<t_{n-k-1\;\alpha} \\
c) & H_1: & \beta_j{\neq}\beta_{j,0} & T_j^*<t_{n-k-1\;\alpha/2}\text{ or }T_j^*>t_{n-k-1\;1-\alpha/2}
\end{array}
$$

In what follows, we focus on simple linear regression. In that case : $c_{0,0}=\frac{\overline{x^2}}{S_{xx}}$ $\quad c_{1,1}=\frac{1}{S_{xx}}$. The confidence intervals for β_1 and β_0 at level α are provided respectively by

$$\beta_1\in\left[\widehat{\beta}_1-t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{S_{xx}^{-1}}\;;\;\widehat{\beta}_1+t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{S_{xx}^{-1}}\right]$$
$$\beta_0\in\left[\widehat{\beta}_0-t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\overline{x^2}S_{xx}^{-1}}\;;\;\widehat{\beta}_0+t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\overline{x^2}S_{xx}^{-1}}\right]$$

The prediction interval for Y_0 at level α is provided by

$$\left[\widehat{y}_0-S_{pred}\,t_{n-2\,;\,1-\frac{\alpha}{2}}\;;\;\widehat{y}_0+S_{pred}\,t_{n-2\,;\,1-\frac{\alpha}{2}}\right]\;,$$

$$S^2_{pred}=\widehat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)$$