Statistical Modelling and Analysis: Formula Sheet

Summary Statistics

For data x_1,x_2,...,x_n

$$\text{Sample mean } \bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \quad \text{Sample variance } s^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right).$$

Discrete Random Variables

		Range	p.m.f.	Mean	Variance
Binomial	$Bin(n,p)$	$0\leq x\leq n$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson	$Po(\lambda)$	$0\leq x<\infty$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
Geometric	$Geo(p)$	$1\leq x<\infty$	$p(1-p)^{x-1}$	$1/p$	$(1-p)/p^2$

Continuous Random Variables

		Range	p.d.f.	Mean	Variance
Uniform	$U(a,b)$	$a\leq x\leq b$	$1/(b-a)$	$(a+b)/2$	$(b-a)^2/12$
Exponential	$Exp(\lambda)$	$0<x<\infty$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
Gamma	$\Gamma(\alpha,\beta)$	$0<x<\infty$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	α/β	α/β^2
Beta	$B(\alpha,\beta)$	$0\leq x\leq 1$	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\alpha/(\alpha+\beta)$	$\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$
Normal	$N(\mu,\sigma^2)$	$-\infty<x<\infty$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2

One Sample Two-sided Confidence Intervals, $(100-\alpha)\%$

Mean (known σ^2)	$x_1,...,x_n$	$N(\mu,\sigma^2)$	$\bar{x}\pm z\times\sigma/\sqrt{n}$	where $\mathbb{P}(Z>z)=\alpha/2$ with $Z\sim N(0,1)$.
Mean (unknown σ^2)	$x_1,...,x_n$	$N(\mu,\sigma^2)$	$\bar{x}\pm t\times s/\sqrt{n}$	where $\mathbb{P}(X>t)=\alpha/2$ with $X\sim t_{n-1}$.
Mean (large sample)	$x_1,...,x_n$	Unknown	$\bar{x}\pm z\times s/\sqrt{n}$	where $\mathbb{P}(Z>z)=\alpha/2$ with $Z\sim N(0,1)$.
Variance	$x_1,...,x_n$	$N(\mu,\sigma^2)$	$\left(\frac{s^2(n-1)}{b},\frac{s^2(n-1)}{a}\right)$	where $\mathbb{P}(\chi^2_{n-1}<a)=\alpha/2$ $\mathbb{P}(\chi^2_{n-1}>b)=\alpha/2$.
Proportion	x	$Bin(n,p)$	$\hat{p}\pm z\times\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	where $\mathbb{P}(Z>z)=\alpha/2$ with $Z\sim N(0,1)$, $\hat{p}=x/n$.
Poisson Mean	$x_1,...,x_n$	$Po(\lambda)$	$\bar{x}\pm z\times\sqrt{\bar{x}/n}$	where $\mathbb{P}(Z>z)=\alpha/2$ with $Z\sim N(0,1)$.

Linear Regression

Summary statistics are:

$$S_{xx}=\sum x_i^2-\frac{(\sum x_i)^2}{n} \qquad S_{xy}=\sum x_i y_i-\frac{(\sum x_i)(\sum y_i)}{n} \qquad S_{yy}=\sum y_i^2-\frac{(\sum y_i)^2}{n}$$

The least squares estimates of α and β in the regression model $y=\alpha+\beta x$ are

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}} \qquad \hat{\alpha}=\frac{1}{n}\left(\sum y_i-\hat{\beta}\sum x_i\right)$$

The residual (error) mean square can be calculated from:

$$s^2=\frac{1}{n-2}\left(S_{yy}-\frac{S_{xy}^2}{S_{xx}}\right)$$

The standard error of the estimate $\hat{\beta}$ is: $\dfrac{s}{\sqrt{S_{xx}}}$ and of the estimate $\hat{\alpha}$ is: $s\sqrt{\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}}$

The $(100-\gamma)\%$ confidence intervals for α and β are given by $\hat{\alpha}\pm t\times ese(\hat{\alpha})$ and $\hat{\beta}\pm t\times ese(\hat{\beta})$ respectively, where $\mathbb{P}(X>t)=\gamma/2$ with $X\sim t_{n-2}$.

The prediction for a given value of the predictor x is: $\hat{y}=\hat{\alpha}+\hat{\beta}x$ and the $(100-\gamma)\%$ confidence interval is given by

$$\hat{y}\pm t\,s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$$

where $\mathbb{P}(X>t)=\gamma/2$ with $X\sim t_{n-2}$.

Pearson's correlation coefficient:

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$