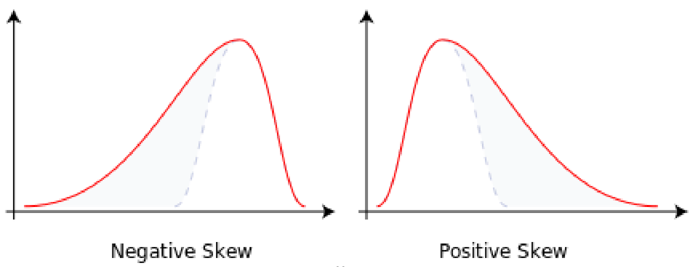
STA2601 Applied Satistics II

Chapter 1

Integral is the distribution function, or cumulative probability function, $F_X(x)$
Derivative is the probability function, or probability density function $f_x(x)$

$Z = \dfrac{X - \mu}{\sigma}$



Exponential distribution: $\frac{1}{k}e^{-\frac{x}{k}}$ parameter is k.

$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$

$\sigma^2 = \mathrm{E}(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx$

Var$(X+Y) =$ Var$(X) +$ Var$(Y) + 2$Cov(X,Y).

$$\begin{aligned} Cov(X,Y) &= E\left[(X - E[X])\,(Y - E[Y])\right] \\ &= E[XY] - E[X]\,E[Y]. \end{aligned}$$

Study Unit 2

→ Unbiased estimator $\mathrm{E}(T) = \theta$
→ $Var(cX) = c^2 Var(X)$
→ Most efficient estimator has the smallest variance.

Least squares estimation:

1) $Q(\theta_1, \dots, \theta_k) = \sum_{i=1}^{n} (X_i - \mathrm{E}(X_i))^2$
2) Replace $\mathrm{E}(X_i)$ with given θ
3) Take partials $\dfrac{\partial Q}{\partial \theta_j}$
4) Set partials $= 0$ and solve for each $\hat{\theta}_i$ then solve simultaneousness equations.

Maximum likelihood estimation: (MLE)

1) Get likelihood function $L(\theta) = f_X(X_1, \theta) f_X(X_2, \theta) \dots f_X(X_n, \theta) = \prod_{i=1}^{n} f_X(X_i, \theta)$
2) Take log of likelihood function if needed
3) Take partials $\dfrac{\partial Q}{\partial \theta_j}$
4) Set partials $= 0$ and solve for $\hat{\theta}$ **! Remember hat** $\hat{}$

		Decision based on the data	
		Do not reject H_0	Reject H_0
The true state of nature	H_0 is true	Good decision	Type I error α
	H_1 is true	Type II error β	Good decision

Significance level $\alpha = P(H_0$ is rejected$|H_0$ is true$)$
Power of the test $1-\beta = P($ not rejecting $H_0|H_1$ is true$)$
Confidence level $1-\alpha$

- If this p-value is very small, $\bar{x}$ is said to be highly significant (usually if $p \ll \alpha$).
- If the p-value is fairly small, $\bar{x}$ is said to be significant (usually if $p < \alpha$).
- If the p-value is large, $\bar{x}$ is said to be not significant (usually if $p > \alpha$).

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$
$$Var(X) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2$$
$$\bar{X} \sim n\left(\theta, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\sqrt{n}(\bar{X}-\theta)}{\sigma}$$

Theorem 1.4
Let U and V be independent variates such that $U \sim n\,(0;\ 1)$ and $V \sim \chi^2_d$ and let

$$T = \frac{U}{\sqrt{V/d}}.\ \text{Then } T \sim t_d.$$

$$\mathrm{E}(X_i^2) = \theta + \mu^2 \quad \mathrm{E}(\bar{X}^2) = \tfrac{\theta}{n} + \mu^2$$

$$\begin{aligned} Var(X_i) &= E(X_i^2) - \mu^2 \\ \Longrightarrow E(X_i^2) &= Var(X_i) + \mu^2 \\ &= \theta + \mu^2 \end{aligned}$$

$$\begin{aligned} Var(\overline{X}) &= E(\overline{X}^2) - \mu^2 \\ \Longrightarrow E(\overline{X}^2) &= Var(\overline{X}) + \mu^2 \\ &= \frac{\theta}{n} + \mu^2 \end{aligned}$$

Study Unit 4

Van der Waerden's formula: $\dfrac{r_i}{n+1}$

χ^2 Goodness-of-fit test:

$$Y^2 = \sum_{i=1}^{k}\frac{(N_i - n\pi_i)^2}{n\pi_i}$$
k number of intervals
N_i is then number of items falling into the interval calculated from $X_i = \sigma Z + \mu$
$n\pi_i$ is interval weighting X number of observations.

If $\mathbf{n\pi_i} < 5$ then pool two or more cells

$Y^2 \sim \chi^2_{k-1}$ if distribution is fully specified, otherwise
$Y^2 \sim \chi^2_{k-1-r}$ where r is number of unknown parameters.

Hypotheses:
H_0: The sample comes from a $n(\mu, \sigma^2)$ distribution
H_1: The sample does not come from a $n(\mu, \sigma^2)$ distribution
Reject null hypothesis if $Y^2 > \chi^2_{k-1-r}$
Compute the MLEs based on the ungrouped data. Compute Y^2 as before.

If $Y^2 < \chi^2_{\alpha;k-r-1}$: do not reject H_0

If $Y^2 > \chi^2_{\alpha;k-1}$: reject H_0

If $\chi^2_{\alpha;k-r-1} < Y^2 < \chi^2_{\alpha;k-1}$: decision uncertain

μ known:	$\hat{\sigma}^2 = \frac{1}{n}\Sigma(X_i - \mu)^2$
σ^2 known:	$\hat{\mu} = \frac{1}{n}\Sigma X_i = \overline{X}$
μ and σ^2 unknown:	$\hat{\mu} = \frac{1}{n}\Sigma X_i = \overline{X}$
	$\hat{\sigma}^2 = \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2$

Use above when testing grouped data from a normal distribution because MLE's are difficult to compute.

Method of moments test for normality

Must perform two hypotheses tests

Skewness

$$\beta_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}}$$

Hypotheses:
H_0: $B_1 = 0$ i.e. symmetric
→ One-sided H_1: $B_1 < 0$ Reject if $B_1 < -TableValue$ or H_1: $B_1 > 0$ Reject if $B_1 > TableValue$. 5% level
→ Two-sided H_1: $B_1 \neq 0$ Reject if $|B_1| > TableValue$. 10% level

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

Hypotheses:
H_0: $B_2 = 3$ i.e. from normal distribution
$B_2 > 3$ leptokurtic
$B_2 < 3$ platykurtic

If the alternative is $\beta_2 < 3$, reject H_0 at the 5% level if $B_2 <$ lower 5% point in table B.

If the alternative is $\beta_2 > 3$, reject H_0 at the 5% level if $B_2 >$ upper 5% point in table B.

If the alternative is $\beta_2 \neq 3$, reject H_0 at the 10% level if $B_2 <$ lower 5% point or if $B_2 >$ upper 5% point in table B.

Use **standardised mean deviation** to test kurtosis if sample size is smaller than 50, statistic A
Same hypotheses as B_2, 4th moment test.

$\sum(X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$

Study Unit 5

Contingency Table Analysis

Fixed Grand total:

> The null hypothesis of independence is
>
> $$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}; \quad i = 1, ..., h; \quad j = 1, ..., k.$$

Fixed row or column totals:

> The null hypothesis of independence is that the probability of falling into category i is the same for all k populations:
>
> $$H_0 : \pi_{i1} = \pi_{i2} = ... = \pi_{ik} \quad \text{for} \quad i = 1, ..., h.$$

Test Statistic:

$$e_{ij} = \frac{N_{i.}N_{.j}}{N_{..}}$$

$$Y^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(N_{ij} - e_{ij})^2}{e_{ij}}.$$

Under the null hypothesis the distribution of Y^2 is approximately that of χ^2 with $(h-1)(k-1)$ degrees of freedom.

Two sided test only i.e. $|Y^2|$.

Don't forget to pool $\mathbf{e_{ij} < 5}$

Exact test for 2x2 table

Discreet Distribution

Find smallest row/column as k then smallest other as n and x is the cell frequency.

$H0$: There is no association between attribute A and Attribute B

$H1$: Have to figure it out. (Is the value of x unusually to large or small to ascribe to chance.

One sided:

Two sided test $\frac{\alpha}{2}$ reject H_0 if x is a rare event.

Correlation

$$R=\frac{\Sigma\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\Sigma\left(X_i-\overline{X}\right)^2\Sigma\left(Y_i-\overline{Y}\right)^2}}. \qquad R=\frac{\Sigma X_iY_i-\dfrac{(\Sigma x_i)(\Sigma Y_i)}{n}}{\sqrt{\left(\Sigma X_i^2-\dfrac{(\Sigma X_i)^2}{n}\right)\left(\Sigma Y_i^2-\dfrac{(\Sigma Y_i)^2}{n}\right)}}.$$

In using R X and Y should follow conditions for bivariate normality:

1) Marginal normality
2) Linearity

Testing for zero correlation $\rho = 0$

$H_0 : \rho = 0$ against the alternatives

$H_1 : \rho < 0$ or

$H_1 : \rho > 0$ or

$H_1 : \rho \neq 0.$

$$T=\frac{\sqrt{n-2}R}{\sqrt{1-R^2}}$$

Student's t-distribution ~ T_{n-2}

$$\Sigma\left(X_i-\overline{X}\right)^2=\Sigma X_i^2-\left(\Sigma X_i\right)^2/n$$
$$\Sigma\left(Y_i-\overline{Y}\right)^2=\Sigma Y_i^2-\left(\Sigma Y_i\right)^2/n$$
$$\Sigma\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)=\Sigma X_iY_i-\left(\Sigma X_i\right)\left(\Sigma Y_i\right)/n.$$

Correlation between X and Y is the same for any linear transformation of X or Y with positive coefficients.

Fisher's Z-transformation for $\rho \neq 0$

Theorem 5.5

Let R be the sample correlation coefficient of a random sample from a bivariate normal distribution.

Let $U=\frac{1}{2}\log_e\frac{1+R}{1-R}$ and $\eta=\frac{1}{2}\log_e\frac{1+\rho}{1-\rho}$.

Then, for large samples, $Z=\sqrt{n-3}\,(U-\eta)$ is approximately a $n\,(0;\,1)$ variate.

$H_0 : \rho = \rho_0$ against
$H_1 : \rho > \rho_0$ or
$H_1 : \rho < \rho_0$ or
$H_1 : \rho \neq \rho_0$ we compute

Confidence interval for ρ

$$\rho=\frac{e^{\eta}-e^{-\eta}}{e^{\eta}+e^{-\eta}}=\tanh(\eta)$$

Do algebraic transformation on $1-\alpha=P(a<Z<b)$ with $Z=$(fisher transform) and then use to get interval OR use table X inversely by linear interpolation.

Equality of two R

$H_0 : \rho_1 = \rho_2$

If $\rho_1 = \rho_2$, that is $\eta_1 = \eta_2$, then $Z = \dfrac{U_1 - U_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$ is approximately $n\,(0;\ 1)$.

Study Unit 6

Single sample

Hypothesis testing

We want to test the null hypothesis $\mathbf{H_0}$: $\boldsymbol{\sigma^2} = \mathbf{c}$.

(a) μ **known**

The procedure is based on the statistic $U = \sum_{i=1}^{n} (X_i - \mu)^2 / c$ which, if H_0 is true, is a χ^2_n variate.

If $\Sigma (X_i - \mu)^2$ is small, it is an indication that σ^2 is small and vice versa. We reject $H_0 : \sigma^2 = c$ against the alternatives

(i) $H_1 : \sigma^2 \neq c$ if $U < \chi^2_{1-\frac{1}{2}\alpha;n}$ or $U > \chi^2_{\frac{1}{2}\alpha;n}$

(ii) $H_1 : \sigma^2 < c$ if $U < \chi^2_{1-\alpha;n}$

(iii) $H_1 : \sigma^2 > c$ if $U > \chi^2_{\alpha;n}$.

(b) μ **unknown**

The procedure is based on the statistic $U = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / c$ which, if H_0 is true, is a χ^2_{n-1} variate.

We reject $H_0 : \sigma^2 = c$ against the alternatives

(i) $H_1 : \sigma^2 \neq c$ if $U < \chi^2_{1-\frac{1}{2}\alpha;n-1}$ or $U > \chi^2_{\frac{1}{2}\alpha;n-1}$

(ii) $H_1 : \sigma^2 < c$ if $U < \chi^2_{1-\alpha;n-1}$

(iii) $H_1 : \sigma^2 > c$ if $U > \chi^2_{\alpha;n-1}$.

Two independent samples

$$F = \frac{\left(\chi^2_{n_1-1}\right) / (n_1 - 1)}{\left(\chi^2_{n_2-1}\right) / (n_2 - 1)}$$

yields test statistic

$$F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \sim F_{n_1-1;n_2-1}.$$

Hypotheses:
$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = c$
$H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq c$
$H_1 : \frac{\sigma_2^2}{\sigma_1^2} < c$
$H_1 : \frac{\sigma_2^2}{\sigma_1^2} > c$

$F \sim F_{f;g}$ then $\frac{1}{F} \sim F_{g;f}$

Paired observations

$H_0 : \sigma_1^2 = \sigma_2^2$

$$T = \sqrt{n-2}\frac{U_{11} - U_{22}}{2\sqrt{U_{11}U_{22} - U_{12}^2}}$$

has a t_{n-2} distribution provided $H_0 : \sigma_1^2 = \sigma_2^2$ is true.

$U_{11} = \sum(X_{1j} - \bar{X}_1)^2$
$U_{22} = \sum(X_{2j} - \bar{X}_2)^2$
$U_{12} = \sum(X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)$
Multiply out to get calculator formulas.

More than two independent samples

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ against the alternative $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$.

Test Statistic

$$U = \max_i S_i^2 / \min_i S_i^2$$

$n-1$ degrees of freedom, $n =$ size of samples.
$k =$ number of sample variances.

Study Unit 7

One Sample Problem

(a) $\overline{X}$ is a $n(\mu;\ \sigma^2/n)$ variate, that is $\sqrt{n}\left(\overline{X} - \mu\right)/\sigma$ is a $n(0;\ 1)$ variate;

known variance

(b) $(n-1)S^2/\sigma^2$ is a χ^2_{n-1} variate;

(c) $\overline{X}$ and S^2 are independent;

unknown variance

(d) $T = \sqrt{n}\left(\overline{X} - \mu\right)/S$ is a t_{n-1} variate.

95% Tolerance Interval: 95% confidence that at least 90% of the population data lie between x and y.

Power of the test.

Given that Z_0 is standard normal Solve for $\bar{X}$ at significance level.
Then use β = P(type II error) = P(H_0 is not rejected|H_1 is true)
and standardise using do not reject signs and get probability
Then power of test is 1-β

Two-sample problem, independent samples. same variance

Theorem 7.2

$$T = \frac{U}{\sqrt{\dfrac{W}{(n_1+n_2-2)}}} \sim t_{n_1+n_2-2}$$

where
$$T = \frac{\left[(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)\right] / \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}{\sqrt{\left[(n_1-1)S_1^2 + (n_2-1)S_2^2\right] / (n_1+n_2-2)}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

and
$$S_p^2 = \frac{\left[(n_1-1)S_1^2 + (n_2-1)S_2^2\right]}{(n_1+n_2-2)}$$

$$= \frac{\left[\sum_{j=1}^{n_1} (X_{1j} - \overline{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \overline{X}_2)^2\right]}{(n_1+n_2-2)}$$

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

$H_1 : \mu_1 - \mu_2 < 0$

$H_1 : \mu_1 - \mu_2 > 0$

Paired observations

Same as one sample problem, by transforming by subtraction, then use $T = \dfrac{\sqrt{n}(\bar{Y} - \mu)}{S_Y}$

Independent samples with unequal variances

We want to test $H_0 : \mu_1 - \mu_2 = c$ (with c specified) against
$H_1 : \mu_1 - \mu_2 \neq c$ or
$H_1 : \mu_1 - \mu_2 < c$ or
$H_1 : \mu_1 - \mu_2 > c$.

We use the statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2 - c}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

to get degrees of freedom, use formula and the interpolate table III.

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2\left(n_1-1\right)} + \frac{S_2^4}{n_2^2\left(n_2-1\right)}}.$$

More than two independent samples One-way ANOVA

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2 \quad MSE = \frac{SSE}{(kn-k)}$$

$$SSTr = n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2 = \sigma^2 U \qquad MSTr = \frac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}{k-1} = \frac{\sigma^2 U}{(k-1)}$$

Theorem 7.7

Let $F = \frac{MSTr}{MSE} = \frac{\sigma^2 U/(k-1)}{\sigma^2 V/(kn-k)} = \frac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2/(kn-k)}.$

Then $F \sim F_{k-1;kn-k}$ if $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is true.

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	0.675	3	0.225	5.625
Error	0.640	16	0.04	
Total	1.315	19		

Multiple comparisons for ANOVA

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{S\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{n}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}S}$$

and reject $H_0(p;q) : \mu_p = \mu_q$ in favour of

$H_1(p;q) : \mu_p \neq \mu_q$ if $|T_{pq}|$ exceeds a critical value.

$|T_{pq}| > \sqrt{(k-1)\,F_{\alpha;k-1;kn-k}}.$

Study Unit 8

The maximum likelihood estimators for $\beta_0,\ \beta_1$ and σ^2 are

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

$$\hat{\beta}_1 = \frac{\Sigma Y_i\left(X_i - \overline{X}\right)}{\Sigma\left(X_i - \overline{X}\right)^2}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2.$$

$$\hat{\beta}_1 = \frac{\sum(X - r\bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

Theorem 8.1

$\hat{\beta}_0$ and $\hat{\beta}_1$ are jointly normally distributed with

$E\left(\hat{\beta}_0\right) = \beta_0;$ $\quad E\left(\hat{\beta}_1\right) = \beta_1;$

$Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{d^2}\right);$ $\quad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{d^2};$

$Cov\left(\hat{\beta}_0;\ \hat{\beta}_1\right) = \frac{-\sigma^2\overline{X}}{d^2}$ $\quad$ with $d^2 = \Sigma\left(X_i - \overline{X}\right)^2.$

Theorem 8.2

$$\frac{\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\sigma^2} \sim \chi^2_{n-2}$$

and is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

We call $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ the *predictions* and $Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ the residuals, that is the difference between the observations and predictions.

Inference on the coefficients

Theorem 8.2

$T_0 = \dfrac{\hat{\beta}_0 - \beta_0}{S\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{d^2}}}$ is a t_{n-2} variate.

Theorem 8.3

$T_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\dfrac{S}{d}}$ is a t_{n-2} variate.

Variance around regression line:

$$S^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{(n-2)}.$$

Confidence interval for B_0

$$\widehat{\beta}_0 \pm t_{0.025;\,13} SE(\widehat{\beta}_0)$$

Confidence interval for B_1

$$\widehat{\beta}_1 \pm t_{0.025;\,13} SE(\widehat{\beta}_1)$$

Inference on the regression line

$$\begin{aligned} E\left[\hat{Y}(X)\right] &= E\left(\hat{\beta}_0\right) + XE\left(\hat{\beta}_1\right) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

$$\begin{aligned} Var\left[\hat{Y}(X)\right] &= Var\left(\hat{\beta}_0\right) + X^2 Var\left(\hat{\beta}_1\right) + 2Xcov\left(\hat{\beta}_0;\ \hat{\beta}_1\right) \\ &= \sigma^2\left[\left(\frac{1}{n} + \frac{\overline{X}^2}{d^2} + \frac{X^2}{d^2} - \frac{2X\overline{X}}{d^2}\right)\right] \\ &= \sigma^2\left[\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}\right] \end{aligned}$$

$$\left(\hat{\beta}_0 + \hat{\beta}_1 X\right) \pm t_{\frac{\alpha}{2};n-2} S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}}$$

Relationship between test for correlation and regression

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

$$S^2 = \frac{1}{n-2}\Sigma\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2.$$

$Var(\sum b_i Y_i) = \sigma^2 \sum b_i^2$

to standardise: $\dfrac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(x)}}$