NGUYEN XUAN BINH Id: 887799
Statistical Inference

Population: all of items
Samples: Subset of pop
Observation: ele of sample

median: 0.5 quartile
quantile of an obs:
ith obs = $q(n+1)$
where $q \in [0,1]$
0.25 & 0.75 quantile are 1st & 3rd of
- mid hinge: $\frac{1}{2}(x_{(1st\,q)} + x_{(3rd\,q)})$
- mode: highest frequency
- MAD (median abs devi) is median of $|x_n - m|$
n: 1 → n. Robust against outl
- Range: $[x_{min}, x_{max}]$
- IQT range $[y_{1st\,q}, y_{3rd\,q}]$
- Length of range: $x_{max} - x_{min}$

▫ Four moments of distribution
mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
std = $\sqrt{var}$
skewness: $\hat{\gamma} = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$

kurtosis: $\hat{k} = \frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

Skewness: devi from symmetry
Kurtosis: heaviness of tails
▫ cov: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \delta_{xy}$
▫ cor: $\delta_{xy}/(\delta_x \cdot \delta_y)$

confidence interval & Ho testing
if 95% CI for param θ are computed from 100 indpd samples, 95 of them will contain the true θ but we do not know which
- Bootstrap methods: drawing n obs from data with replacement & times repeat procedure to many
B: num of bootstraps
⇒ $100(1-\alpha)\%$ CI is from $B_{x(\alpha/2)}$ to $B_{(1-\alpha/2)}$
- Exact CI, ND(normal dist)
$(\bar{x} - t_{n-1}^{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}^{\alpha/2}\frac{s}{\sqrt{n}})$

- Hypothesis testing
Null: H_0, alt: H_1
p-value < α: stats significant ⇒ ✗ H_0, ✓ H_1
p-value ≥ α: stats insignificant ⇒ ✓ H_0, ✗ H_1
Type I error: reject true H_0. Prob ≤ α
Type II error: accept false H_0: Prob: β and test power = $1-\beta$
- Test statistic is p-value random var from sample data & used in hypothesis test
- Stats H_0 test: select stats model, state H_0, H_1 → Select test statistic → pick sample → calculate test stat value from sample → calculate p-value from test stat → reject/accept H_0

if X not iid ⇒ 1 samp test Can't be used

▫ One sample t-test
- ASS: i.i.d sample from ND
- Hypo: $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$
- t-test stat
$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$
⇒ $t \approx 0 \Rightarrow H_0$ ✓
|t| large ⇒ H_1 ✓ (Welch)

▫ Two sample t-test^v
- ASS: X iid from ND$_X$, Y iid from ND$_Y$. X and Y are independent
- Hypo: $H_0: \mu_X = \mu_Y$
$t \approx 0 \Rightarrow H_0$ ✓, |t| big ⇒ H_1 ✓

▫ Paired t-test
- ASS: a pair X_1 and X_2 can be from same data and dependent. $X_1 - X_2$ follows ND. $X_1 - X_2 = d$
- Hypo: $H_0: \mu_d = 0$
Apply one samp t-test on d ⇒ $t \approx 0 \Rightarrow H_0$ ✓

▫ Variance test
- ASS: i.i.d from ND
- Hypo $H_0: \sigma^2 = \sigma_0^2$
- χ^2 test stat
$\chi^2 = (n-1)s^2/\sigma_0^2$
$E[\chi^2] = n-1 \Rightarrow H_0$ ✓

▫ variance comparison test
- ASS: X iid ND$_X$, Y iid ND$_Y$. X & Y are independent
- Hypo $H_0: \sigma_x^2 = \sigma_y^2$
- F test stats: S_x^2/S_y^2
$E[F] \approx 1 \Rightarrow H_0$ ✓

Non-parametric test
- Parametric: defined by sets of params & follow a certain kind of distribution
- Non-parametric: does not assume a family of distri

▫ One-sample sign test
- ASS: X iid from continuous distribution
- Hypo H_0: med = m_0
- test $S = n(x_i > m_0)$
⇒ $S \approx 1/2 * n \Rightarrow H_0$ ✓
Standard: $Z = \frac{S - n/2}{\sqrt{n/4}}$

▫ Paired sign test
- ASS: X_1, X_2 can be dependent. $d = X_1 - X_2$ is continuous
- Hypo H_0: med$_d$ = 0
Apply one samp sign test to d

▫ 1 samp Signed rank/ Wilcoxan test
- ASS: X iid from continuous, symmetric dist
- Hypo H_0: med = m_0
order $|d_i| = |x_i - m_0|$ from small to largest
⇒ r_i = rank($|d_i|$) × sign of $(x_i - m_0)$
- test stat: $W = \sum r_i$
$W = \frac{n(n+1)}{4} \Rightarrow H_0$ ✓

▫ Paired signed rank test
- ASS: X_1, X_2 can be dependent, $d = X_1 - X_2$ has symmetric dist
- Hypo H_0: med$_d$ = 0
Apply 1 samp signed test to d

▫ Conclusion: 1 samp compare θ to const. Paired samps compare θ_1 & θ_2
- If ND → t-test
- If symmetric continuous → signed rank test
- If continuous → sign test

Bernoulli dist: μ
$P(x=1) = p$, $P(x=0) = 1-p$
$E(x) = p$, $var(x) = p(1-p)$
For large n, $100(1-\alpha)\%$ CI for success prob p is $[\hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + \dots]$

▫ 1 samp proportion test
- ASS: X iid from Bernoulli
- Hypo H_0: $p = p_0$
- test: $C = \sum x_i$
⇒ $E[C] = np_0$ and $var(C) = np_0(1-p_0)$ ⇒ H_0 ✓

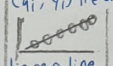
▫ 2-sample proportion test
- ASS: X iid, Y iid from Bernoulli. X & Y independent
- Hypo H_0: $p_X = p_Y$
$\hat{p} = (n\hat{p}_X + m\hat{p}_Y)/(n+m)$
$Z = (\hat{p}_X - \hat{p}_Y)/\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}$
- n ↑ ⇒ Z ↑ ⇒ p-value ↓

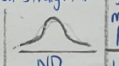
* Normality testing
- ASS: X iid from a rand var x of standard
- Hypo H_0: x follows ND

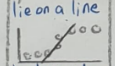
▫ Bowman-Shenton test
test: $BS = n\left(\frac{\hat{\gamma}^2}{6} + \frac{\hat{k}^2}{24}\right)$
$E[BS] = 2 \Rightarrow H_0$ ✓

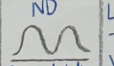
▫ Shapiro-Wilk test
- Hypo H_0: x follows ND
test stat: squared value of Pearson sample correlation
large test value ⇒ H_0 ✓
Both Bowman & Shapiro test requires large sample

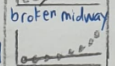
▫ Rank plot/QQ plot
Let $y_1 \leq y_2 \leq \dots \leq y_n$ be datapoints $x_1, x_2 \dots x_n$ ordered from small → largest
- Let $q_i = i/(n+1)$ quantile from standard ND($\mu = 0$, $\delta = 1$) and plot the pairs (q_i, y_i), $i = 1, 2, \dots n$
- if x_i follows ND ⇒ (q_i, y_i) lie on straight line

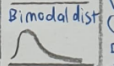

lie on a line

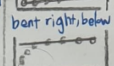

ND

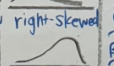

broken midway

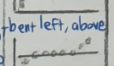

Bimodal dist

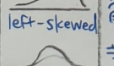

bent right, below

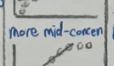

right-skewed

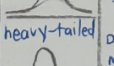

bent left, above


left-skewed


more mid-concen


heavy-tailed


more spread out


light-tailed

- multinomial distribution is generalized binomial dist with multi vars: $E[x_i] = np_i$

▫ χ^2 goodness-of-fit test
- ASS: X iid from rand x
- Hypo H_0: x follows dist F_X (with or without params)
- Test stat: $\chi^2_g = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$
$\chi^2_g \approx k-1-e \Rightarrow H_0$ ✓
follows χ^2_{k-1-e} dist

▫ χ^2 homogeneity test
- ASS: r samples are independent and all obs in r_i are iid. Also, each r_i sample has n_i obs. Hypo H_0: the samples r_i all come from the same dist F_X
- test: $\chi^2_h = \sum_i^r\sum_j^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
$\chi^2_h \approx (r-1)(k-1) \Rightarrow H_0$ ✓

	1	2	k	sum
1	E_{11}	E_{12}	E_{1k}	n_1
r	E_{r1}	E_{r2}	E_{rk}	n_r
sum	k_1	k_2	k_K	n

$p_J = k_J/n$, $E_{ij} = n_i p_J$

▫ χ^2 test of independence
- ASS: iid rand samp of size n, obs divided into r classes wrt factor A and k classes wrt factor B
- Hypo H_0: var A and B are independent ⇒ test stats $\chi^2_I = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

	1	2	k	sum R_i
1	a_{11}	a_{12}	a_{1k}	R_1
2				
r	a_{r1}	a_{r2}	a_{rk}	R_n
sum	k_1	k_2	k_K	n

$q_i = R_i/n$ $p_J = k_J/n$
⇒ $E_{ij} = nq_ip_j$ for each O_{ij}
$\chi^2_I \approx (r-1)(k-1) \Rightarrow H_0$ ✓
* when to use χ^2_H & χ^2_I check if gender affect vote ⇒ interview n people

	Repub	Demo
Male		
Female		

▫ if chose before hand gender ⇒ no rand any more with marginal freq fixed ⇒ use χ^2_H if they share same dist
▫ if we know beforehand the votes but not gender ⇒ use χ^2_I if gender & vote independent

each r_i sample has n_i obs. Hypo H_0: the samples r_i all come from the same dist F_X
- test: $\chi^2_h = \sum_i^r\sum_j^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
$\chi^2_h \approx (r-1)(k-1) \Rightarrow H_0$ ✓

	1	2	k	sum
1	E_{11}	E_{12}	E_{1k}	n_1
r	E_{r1}	E_{r2}	E_{rk}	n_r
sum	k_1	k_2	k_K	n

$p_J = k_J/n$, $E_{ij} = n_i p_J$

▫ χ^2 test of independence
- ASS: iid rand samp of size n, obs divided into r classes wrt factor A and k classes wrt factor B
- Hypo H_0: var A and B are independent ⇒ test stats $\chi^2_I = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

	1	2	k	sum R_i
1	a_{11}	a_{12}	a_{1k}	R_1
2				
r	a_{r1}	a_{r2}	a_{rk}	R_n
sum	k_1	k_2	k_K	n

$q_i = R_i/n$ $p_J = k_J/n$
⇒ $E_{ij} = nq_ip_j$ for each O_{ij}
$\chi^2_I \approx (r-1)(k-1) \Rightarrow H_0$ ✓
* when to use χ^2_H & χ^2_I check if gender affect vote ⇒ interview n people

	Repub	Demo
Male		
Female		

▫ if chose before hand gender ⇒ no rand any more with marginal freq fixed ⇒ use χ^2_H if they share same dist
▫ if we know beforehand the votes but not gender ⇒ use χ^2_I if gender & vote independent

$var(x) = E[x^2] - E(x)^2$
Pearson corr coeff $\rho(x,y) = \frac{cov(x,y)}{\delta x \delta y} = \frac{\delta xy}{\delta x \delta y}$
$cov(x,y) = E[xy] - E[x]E[y]$
$= E[x - E[x])(y - E[y])]$

▫ Spearman coeff is Pearson coeff calculated from ranks
▫ Pearman coeff measures linear relationship while Spearman coeff measures monotonic relationship [-1, 1]
$P_S(x,y) = 0 \Rightarrow x, y$ independ

Linear regress: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- ASS$_1$: $E(\varepsilon_i) = 0$, $var[\varepsilon_i] = \delta^2$, $\rho(\varepsilon_x, \varepsilon_y) = 0$ ($i \neq j$) and errors are iid

▫ Estimated regress line
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (n_J is group size of Jth group. n is sum of all data pairs)
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \hat{\rho}(x,y)\frac{S_y}{S_x}$
$\Rightarrow \hat{y}_i = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\rho}(x,y)\frac{S_y}{S_x}x_i$
$= \bar{y} + \hat{\rho}(x,y)\frac{S_y}{S_x}(x_i - \bar{x})$

▫ residual $\hat{\varepsilon}_i$ of estimated model: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
(dif between observed & fitted value $\hat{y}_i$) ⇒ $\hat{\varepsilon}_i \downarrow$, obs value is better predicted

error	residual
- unobservable	- observable
- uncalculable	- calculable
- dif between observed data and true value of the popular	- dif between observed data and predicted value of sample

▫ estimated residual mean^2
$\hat{\delta}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

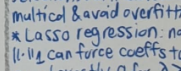
▫ coeff of deter $R^2 \in [0,1]$
$R = 1 - (SSE/SST)$
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ $SSE < SST$ ⇒ model explains better
$SSE = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$

$R^2 = 1 \Leftrightarrow \hat{\varepsilon}_i = 0$ ($\Leftarrow$)
$\hat{\rho}(x,y) = \pm 1$ ($R^2 = \hat{\rho}(x,y)^2$)
$R^2 = 0 \Leftrightarrow \hat{\beta}_1 = 0$ ($\Leftarrow$)
$\hat{\rho}(x,y) = 0$ ($\Leftarrow$) no relation
- R^2 can't be used for variable selection. R^2 just tell how well the model fits the data but not whether coeff estimate & predictions are biased. R^2 can't explain how one var explains another
In multiple linear regress, R^2 can't tell which vars are more important than the others
▫ S lope test
- ASS: $ASS_1 + \varepsilon_i$ follows ND
- Ho: $\beta_1 = \beta_1^0$
▫ Intercep test: Ass above
- Ho: $\beta_0 = \beta_0^0$
▫ $R = 0.66 \Rightarrow$ model explains 2 thirds of variation in the response var ⇒ good fit. R^2 if our lies removed

Multiple linear regression
- ASS: like ASS$_1$ $(i=1 \to n)$
$y_i = \beta_0 + \beta^T x_i + \varepsilon_i$
$\beta = (\beta_1, \dots \beta_p)^T$, $var(\varepsilon_i) = \delta^2$
Estimated model
$\hat{y} = X(X^TX)^{-1}X^Ty$
* Multicollinearity can make regression coeff unstable ⇒ some of vars can be dropped without lost info
▫ Variance inflator-factor
$VIF_k = 1/(1 - R_k^2)$
R_k^2 is R^2 for model where x_{ik} is dependent and remaining predicting vars are used as explain vars
- VIF = 1 (⇐) var x_{ik} uncorrelated with other explain vars and VIF ≥ 10 ⇐) multicollinear is present & some vars should be dropped
▫ Residual mean^2 estimate
$\hat{\delta}^2 = 1/(n-p-1)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
▫ Heteroskedasticity $var(x)$ varies as x varies
▫ Bootstrap in linear model: observation resampling → draw a bootstrap sample size n and fit linear model to it.
▫ Residual resampling: (x_i, y_i^*) where $y_i^* = \hat{y}_i + \hat{\varepsilon}_i^*$ where $\hat{\varepsilon}_i^*$ is bootstrap sample of the residuals of original samp
* Problem: data exhibits non-linear dependency.
⇒ Transform predictors: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \varepsilon_i$ (still multiple regress)
⇒ add interaction terms like $\beta_{12} x_1 x_2$
⇒ kernel regression
* Problem: have several response variables.
⇒ model using multivariate model: $y_i = \beta_0 + \beta^T x_i + \varepsilon_i$
* Problem: many outliers but don't wanna discard
⇒ use robust methods like least absolute value or L_1-regression ($E[\varepsilon_k - 1] = H_0$)
* Problem: datapoints have significant correlations ⇒ use mixed models or time series methods

▫ Variable selection
* Backward selection
- Set $\alpha_0 = 0.05$ (cut-off p-value)
1) Estimate model with all predictors
2) Drop predictor with max p ≥ α_0 and estimate new model
3) Repeat step 2 until all predictors have p-value < α_0
* Forward selection
- Set cutoff p value $\alpha_0 = 0.05$
1) Start model with no predictor
2) for each predictor, check their p-value when added to model & add the one with p-value min < α_0
3) Repeat step 2 until no new predictors with p-value < α_0 can be added
* Drawbacks, miss optimal model as not all combinations are tried ⇒ combi of back & forward selection avoids this
- the more predictors, the more likely type I error
- Cutoff p value might miss almost significant relevant predic
▫ Similar variance ASS tests of ANOVA tested with Barlett
▫ Barlett's test
- ASS: let x_{nJ} iid from ND(μ_J, δ_J^2) of rand x_J. k samples are indpdent
- Ho: $\delta_1^2 = \delta_2^2 = \dots$ ⇒ same
- H_1: $\delta_i^2 \neq \delta_j^2$ for $i \neq j$
- test: $B = Q/H$
$E[B] \approx k-1 \Rightarrow H_0$ ✓
▫ 2 samp rank test: X, Y iid, cont dist. Ho: med x = med y
$E[W] = n(n+m+1)/2 \Rightarrow H_0$ ✓
* Ridge regression: not measure vars importance. The coeffs never reach 0. ⇒ not used for variable selection. But it addresses multicol & avoid overfitting
* Lasso regression: norm $||\cdot||_1$ can force coeffs to equal exactly 0 for $\lambda \geq 0$

Degrees of freedom



⇒ importance: $x_1 > x_2 > x_3 > x_4$

Analysis of variance (ANOVA)
- ASS: x_{nJ} iid from x_J from ND, $J = 1 \to k$. k samp are independent. Groups have equal variance
- Ho: $\mu_1 = \mu_2 = \dots = \mu_K$
- H_1: $\mu_n \neq \mu_J$ for some $n \neq J$
▫ variance between groups
$SSG = \sum_{J=1}^{k} n_J(\bar{x}_J - \bar{x})^2$
▫ variance within group
$SSE = \sum_{J=1}^{k}\sum_{i=1}^{n_J}(x_{iJ} - \bar{x}_J)^2$
▫ F test = $\frac{SSG/(k-1)}{SSE/(n-k)}$ Ho ✓
$E(H_0) = (n-k)/(n+k-2) \Rightarrow$
* Bonferroni correction
To avoid type I error in multiple testing, for each pair μ_J, μ_K, we conduct t-test and reject Ho: $\mu_J = \mu_K$ if their p-value $p < \alpha/C$, C is num of pairs
* Kruskal-Wallis test
ANOVA without ND. Ho: k indpd samps from same dist. med$_0$ = med$_1$ = ... = med n.
if K-W test (reject Ho ⇒ use 2 samp rank test with Bonferroni

* Problem: range on 2 sides
$E[y_i] = \beta_0 + \beta^T x$ not match
⇒ use link function: $g(E[y_i]) = \beta_0 + \beta^T x$
* Prob: too many vars, $p \geq n \Rightarrow X^TX$ singular ⇒ no least squares exist

Nadaraya Watson regression: At any point x: $g(x) = \sum_{i=1}^{n} w_i y_i$

where the weights are calculated as:

$$w_i = \frac{K(x - x_i)}{\sum_{j=1}^{n} K(x - x_j)}$$

$$CV(\lambda) = \frac{1}{n} \sum (y_j - \hat{m}_\lambda^{(-j)} x_j)^2$$

point j left out of fit; estimate $m(\cdot)$ based on other $n-1$ observations

Kernel function: Epachnechnikov kernel: $K_1(u) = \frac{3}{4}(1 - u^2)$ for $-1 \leq u \leq 1$ and 0 outside that interval

Then scaled to bandwidth h with $K_h(u) = K_1(u/h)$, positive for $-h \leq u \leq h$

- Choice of bandwidth

Large bandwidth = averaging many datapoints = very smooth regression function that shows large scale features of the data. Efficiently smooths small errors away

Small bandwidth = averaging few data points = very wild regression function that follows the data very closely: But also retains its error

- Kernel density estimation

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$: the datapoints x_i representing point masses $1/n$ each then doing kernel smoothing to distribute those masses around x_i over some distance by the kernel func. This gives nicer, smoother estimate of unknown density than a histogram

ANOVA:

- group means $\bar{x}_j = 1/n_j \sum_{i=1}^{n_j} x_{ij}$, where n_j is the group size of the jth group
- combined sample mean: $\bar{x} = 1/n \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}$ where $n = \sum_{j=1}^{k} n_j$
- group sum of squares: $SSG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$
- error sum of squares: $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$

{ k: num of groups
j: group index
n_j: size of group j
i: index of ele in group j

- F-test: $F = \frac{SSG/(k-1)}{SSE/(n-k)}$ $H_0 = \frac{n-k}{n-k-2}$

A	4	3	2	4	6	5	4	3	4	4	$\bar{x}_A = 3.9$
B	6	5	4	5	8	5	3	5	5	6	$\bar{x}_B = 5.2$
C	2	3	1	3	2	6	3	5	2	2	$\bar{x}_C = 2.7$
D	6	7	6	5	5	5	6	3	6	7	$\bar{x}_D = 5.6$

group means combined sample mean

$\bar{x} = (\bar{x}_A + \ldots \bar{x}_D)/4 = 4.35$

$SSG = 10(3.9 - 4.35)^2 + 10(5.2 - 4.35)^2 + \ldots = 52.1$

$SSE = (4 - 3.9)^2 + (3 - 3.9)^2 + \ldots (6 - 5.2)^2 + (5 - 5.2)^2 + \ldots$ C + D

$= 55 \Rightarrow F = \frac{(40-4)\,52.1}{(4-1)\,55} = 11.367 \neq \frac{40-4}{40-4-2} = 1.0588$

A B

$\Rightarrow$ At least one group differs

April 14, 2021

P1 In each question, answer TRUE or FALSE. In this problem you do not need to write reasons. 1 point per item for correct answer, maximum amount of points obtainable is 6.

(a) If a distribution is symmetric around its mean, it has a positive skewness coefficient.

(b) The two-sample proportion test can be used even if the two samples have different sizes.

(c) In linear regression, correlation coefficient is the slope of the regression line.

(d) Confidence level indicates how much confidence you have in the model assumptions.

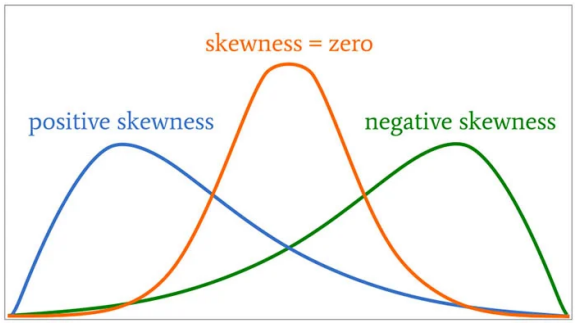
(e) A normal distribution is symmetric around its mean, but there are also other distributions that are symmetric.

(f) Bonferroni correction is a method for using linear regression with nonlinear data.

(g) The chi-squared (χ^2) distribution is skewed to the right.

(h) In hypothesis testing, the null hypothesis is rejected when getting a p-value smaller than the significance level.

(a) False. Symmetric distribution has 0 skewness coefficient



(b) True

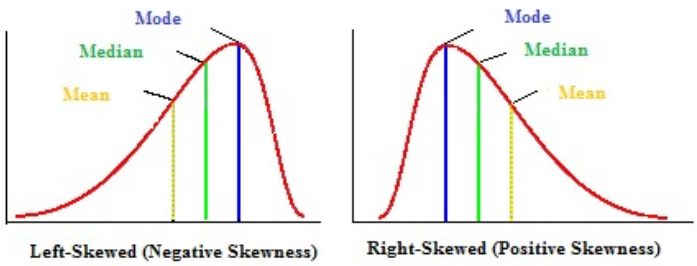
(c) False. Slope of the regression line is called the regression coefficient. Correlation coefficient measures the correlation between two random variables

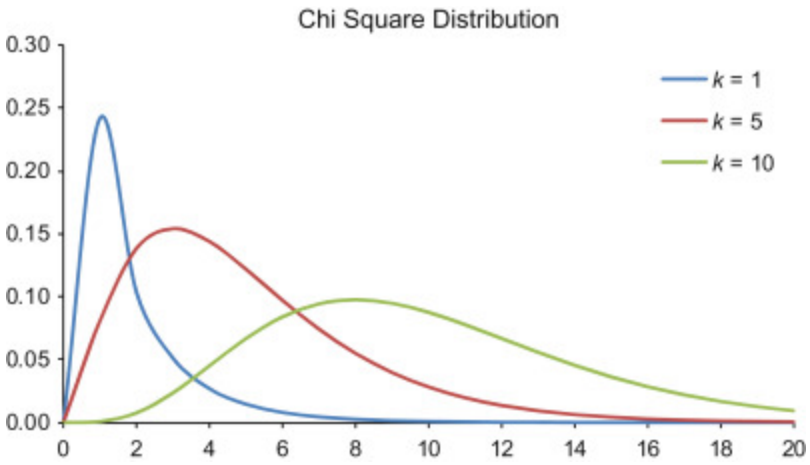
(d) False. A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

(e) True. Example: bimodal distribution

(f) False. It is used in multiple t-test (ANOVA) to prevent type I error, by determining that p-value is smaller than alpha/number of pairs testing

(g) True





(h) True

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give the general definitions of Type I and II errors but instead state what they mean for this specific pair of hypotheses). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name two ways besides Q-Q plot for checking/testing the normality of a sample. **(1p)**

(d) A researcher wants to model her data with Model X that makes a normality assumption. For this, she tests her data for normality and gets a p-value of 0.055 (for the hypotheses given in part a). Based on the p-value, she decides to use Model X. Can the researcher fully trust the results of the model? Explain why or why not. **(1p)**

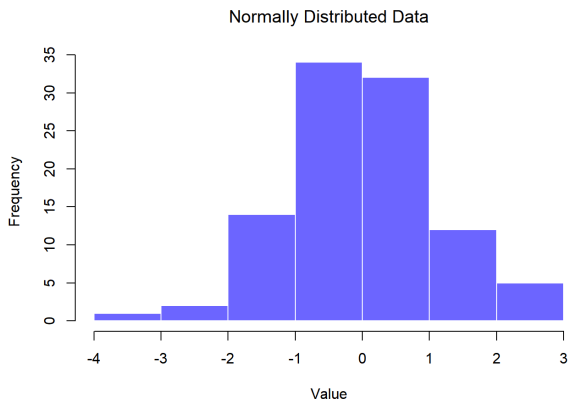
a)

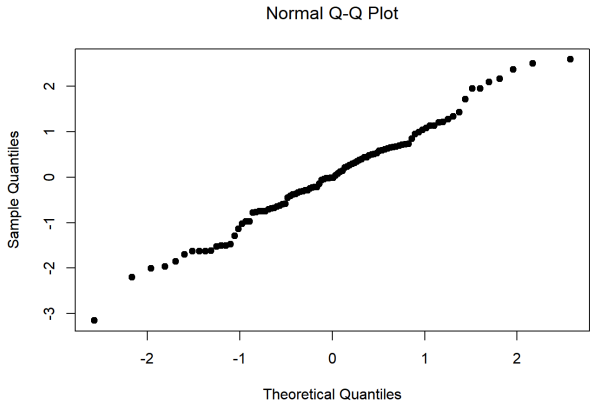
Type I error: reject the true null hypothesis -> Although the true distribution of the samples is normal, it is thought that the samples do not come from a normal distribution

Type II error: accept the false null hypothesis -> Although the true distribution of the samples is not normal, it is thought otherwise that the samples actually come from the normal distribution

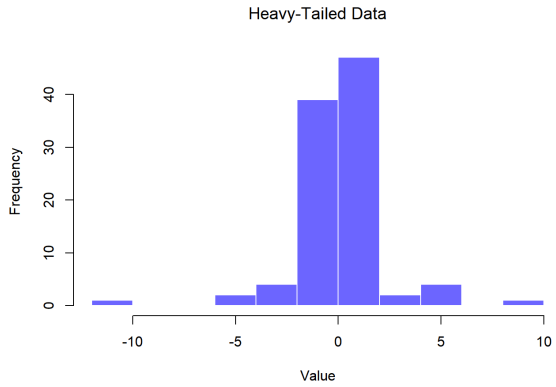
b)

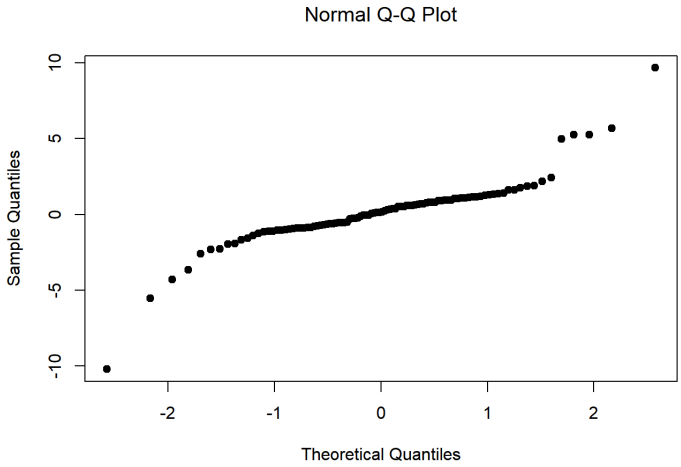
i) Normal distribution: points lie on diagonal line





ii) Not normal distribution





c) Bowman-Shenton and Shapiro-Wilk test.

d) This p value is not statistically significant => Null hypothesis is accepted: the data actually comes from a normal distribution and thus the researcher can trust the results of the model X.

P3 In each of the following scenarios you have an i.i.d. (independent and identically distributed) sample $x_1, \ldots, x_n$ from some distribution F. Describe (in 3–5 sentences, and including also the assumptions that your chosen methods make) how you would investigate the following research questions.

(a) Does the sample come from a distribution with median equal 0? **(2p)**

(b) Does the sample come from the standard normal distribution? **(2p)**

(c) Does the sample come from a distribution whose standard deviation is 5? **(2p)**

a) One sample sign test: This test has assumption the sample comes from a distribution with median is m, in this case m is 0. This test is a non parametric test so that we do not assume the normal distribution. The null hypothesis is H0: m = 0 and alternative hypothesis H1: m # 0.

b) Bowman-Shenton test: This test will test the normality of the sample, it is a function of skewness and kurtosis. Its assumption is the sample is i.i.d which is satisfied. The null hypothesis is H0: The sample is from standard normal

distribution and alternative hypothesis is H1: The sample isn't from standard normal distribution

c) The variance test or Chi-squared test: Assume the sample is normally distributed. The Chi-square test is suitable for testing if the sample is from a distribution whose standard deviation is 5. The null hypothesis is H0: sigma = 5 or sigma^2 = 25 and alternative hypothesis is H1: sigma # 5 or sigma^2 # 25

P4

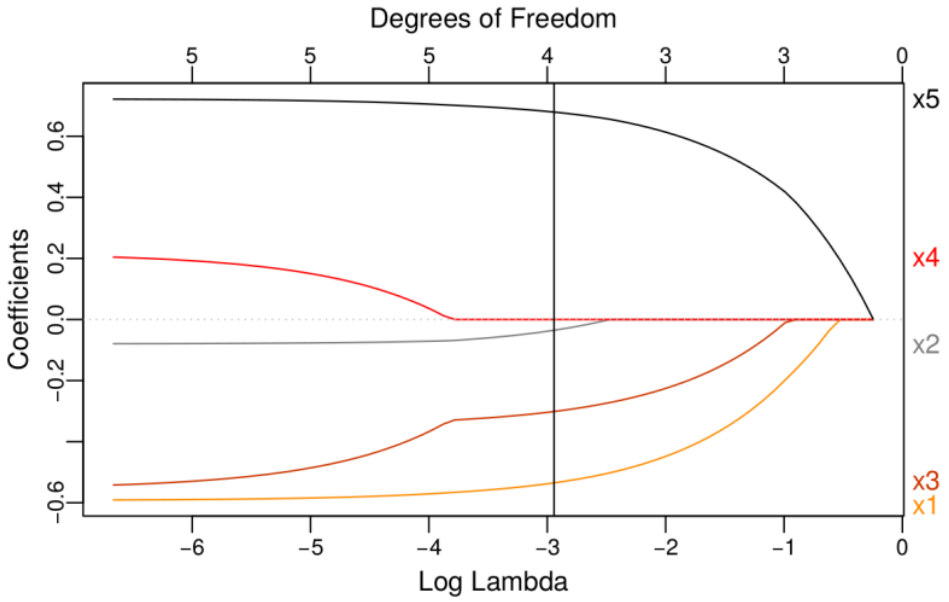
(a) How does backward selection conduct variable selection? List also two of its drawbacks. **(3p)**

(b) Why is simply picking the model which gives the largest value of R^2 not a good idea with respect to variable selection? **(1p)**

(c) The following plot shows the LASSO coefficient profiles in a regression problem with a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. (If the colors are not well visible, note that the order of the profiles from top to bottom in the left end of the plot is `x5`, `x4`, `x2`, `x3`, `x1`). The optimal value of $\log(\lambda)$ given by cross-validation is shown as a vertical line.

(i) Which variable does LASSO hold as the most important one? Which is the second most important? **(1p)**

(ii) Write down (approximately) the estimated coefficients of the five predictors in the model corresponding to the optimal value of $\log(\lambda)$, and explain which variables (if any) have been left out of the model by this stage. **(1p)**



a) The backward selection works by selecting a p-value cutoff α0 (e.g. 0.05) and proceeding as follows:

1 Estimate the model using all predictors.

2 Remove the predictor with the highest p-value greater than or equal to α0 and estimate the new model.

3 Repeat step 2 until all predictors have p-values less than α0.
That is, backward selection begins with a full model and one-by-one removes the variables that are the least important, until we are left with the subset of most important variables
Drawbacks:
1) The absolute p-value cut-off might miss some almost significant predictors which are actually relevant.
2) It is possible to miss the optimal model as not all possible combinations of the predictors are considered during the process. (a combination of the backward and forward selection, stepwise selection, would avoid this)
3) both backward and forward selection get increasingly complex if one allows for interaction terms between the predcitors (e.g. age×sex)

b) One of the most essential limits to using this model is that R-squared cannot be used to determine whether or not the coefficient estimates and predictions are biased. Furthermore, in multiple linear regression, the R-squared can not tell us which regression variable is more important than the other.

c)
(i) Importance of variables from most to least: x5 -> x1 -> x3 -> x2 -> x4
=> most important variable: x5. Second most important variable is x1
(ii) At this stage, the explanatory variable that is left out of the model is x4. The estimated coefficients for the rest variables at optimal value of log(lambda) is:
x1: -0.5 x2: 0.05 x3: -0.3 x5: 0.7

April 8, 2020

P1 Answer either TRUE or FALSE (in this problem, reasons not required). 1 point per item for correct answer, maximum amount of points obtainable is 6.

(a) Median is a measure of scatter.

(b) Descriptive statistics aims to draw conclusions about a population based on a sample.

(c) In the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the error term is usually assumed to have expected value one, $\mathbb{E}(\varepsilon_i) = 1$.

(d) If two predictors are highly correlated with each other in linear regression, this can make the coefficient estimates unstable.

(e) In hypothesis testing, the probability of a Type II error is always greater than or equal to the probability of a Type I error.

(f) In hypothesis testing, the null hypothesis is rejected when getting a p-value smaller than the significance level.

(g) LASSO can be used for variable selection.

(h) In bootstrap, the number of observations in each of the bootstrap samples is the same as the number of observations in the original sample.

(a) False. Median is a measure of location
measure of location: mean, median, quantile, mode
measure of scatter: variance, standard deviation, mean absolute deviation MAD, range, skewess, kurtosis
(b) False. It is inferential statistics that aims to draw conclusions
(c) False. The error term is usually assumed to have 0 expected value
(d) True. Multicollinearity makes the model less exact. Consider use forward or backward selections to remove dependent variables
(e) False. Probability of Type I error is at most the alpha level (α). A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). In short, Type I and Type II Error are intertwined: When we make our Type I Error criterion stricter, we increase the likelihood of Type II Error.
https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/type-i-error-type-ii-error-decision/
(f) True. If p-value is larger than significant level alpha then the null hypothesis is accepted
(g) True. In Lasso regression, as alpha term increases, less significant variables will be dropped
(h) True. In Bootstrap method, the sample size of the new sample is the same as the sample size of the original sample. However, each data point can be selected once, multiple times, or not at all.

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give the general definitions of Type I and II errors but instead state what they mean for this specific pair of hypotheses). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name two ways besides Q-Q plot for checking/testing the normality of a sample. **(1p)**

(d) A researcher wants to model her data with Model X that makes a normality assumption. For this, she tests her data for normality and gets a p-value of 0.055 (for the hypotheses given in part a). Based on the p-value, she decides to use Model X. Can the researcher fully trust the results of the model? Explain why or why not. **(1p)**

(Same as next year)

P3 Consider multiple linear regression on a sample of $n = 100$ observations of a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}$. Below are shown the linear regression model summary, variance inflation factors and the diagnostics plot for the model fit.

(a) Describe how one can check whether the assumptions of multiple linear regression are satisfied. Are they satisfied in the current example? **(3p)**

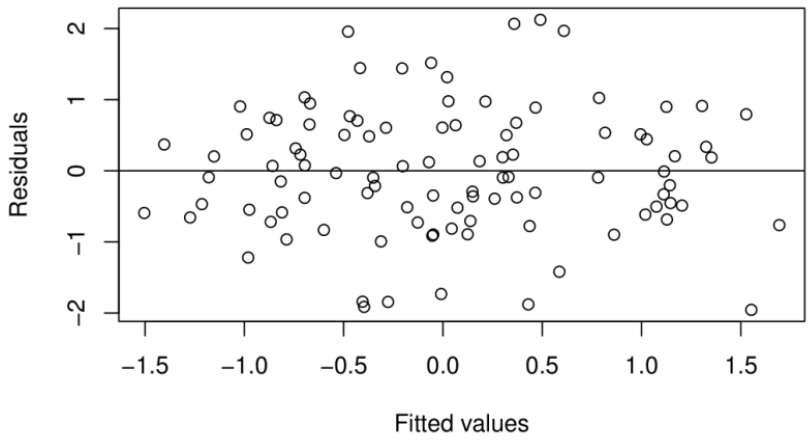
(b) Based on the variance inflation factors, can the estimated coefficients be trusted? Why or why not? Explain what it would mean if these factors are high or low. **(1p)**

(c) Give an interpretation for the estimated coefficient $\hat{\beta}_4 \approx -0.44$. **(1p)**

(d) What does the fitted model predict for the response variable if $x_{i1} = x_{i2} = x_{i3} = 0$ and $x_{i4} = 100$? Give a numerical answer and explain why or why not this prediction can be trusted. **(1p)**

```
## Call:
## lm(formula = y ~ ., data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2158 -0.6744  0.1267  0.6918  1.9987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08357    0.09932  -0.841   0.4022
## x1          -0.59245    0.09081  -6.524 3.42e-09 ***
## x2          -0.07999    0.12061  -0.663   0.5088
## x3          -0.55658    0.21660  -2.570   0.0118 *
## x4           0.21811    0.20657   1.056   0.2937
## x5           0.72269    0.09904   7.297 9.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9499 on 94 degrees of freedom
## Multiple R-squared:  0.5637, Adjusted R-squared:  0.5405
## F-statistic: 24.29 on 5 and 94 DF,  p-value: 1.285e-15

## Variance inflation factors
##       x1       x2       x3       x4       x5
## 1.014837 1.024145 5.757164 5.738664 1.056317
```



a) The assumptions of multiple linear regressions are: E[$error_i$] = 0, var[$error_i$] = sigma^2, cov($error_i$, $error_j$) = 0 for i # j, errors are i.i.d, number of observations are more than the number of explanatory variables (p < n, where p-dimensional xi are non-random)

1) E[$error_i$] = 0: we can see from the residuals plot, the number of data points are symmetric around residual = 0, so this is a true assumption

2) var[$error_i$] = sigma^2 = residual standard error = 0.9499

3) cov($error_i$, $error_j$) = 0: there seems to be no trends in the residuals plot as the points are randomly scattered

4) errors are iid as the points in the residual plots seem to be randomly scattered

5) number of data points is n = 100 and number of explanatory vars is 5 and we have 5 < 100

b) if VIF = 1, var xi is uncorrelated with other explanatory variables and if VIF >= 10, multicollinearity is present and some explanatory vars should be dropped from the model. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. Based on this, the estimated coefficients can be trusted, but cautions should be made towards the coefficients of x3 and x4 where VIF is greater than 5

c) estimated coeff beta_t = -0.44 means that if explanatory var x4 increases by 1 unit, the response var y decreases by -0.44 unit and vice versa => y and x4 are inversely proportional.

d) y_fitted = intercept + beta_1 * x1 + beta_2 * x2 + beta_3 * x3 + beta_4 * x4
= -0.08357 + (-0.59245) * 0 + (-0.07999) * 0 + (-0.55658) * 0 + (0.21811) * 100 = 21.727
This value should not be trusted because it has very great residual

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08357    0.09932  -0.841   0.4022
## x1          -0.59245    0.09081  -6.524 3.42e-09 ***
## x2          -0.07999    0.12061  -0.663   0.5088
## x3          -0.55658    0.21660  -2.570   0.0118 *
## x4           0.21811    0.20657   1.056   0.2937
## x5           0.72269    0.09904   7.297 9.25e-11 ***
```

P4

(a) How can we control the probability of Type I error in hypothesis testing? **(1p)**

(b) Why is the probability of Type II error more difficult to control in hypothesis testing than the probability of Type I error? **(1p)**

(c) When and why should you consider adjusting the significance level (for example, with the Bonferroni correction) in hypothesis testing? **(2p)**

(d) Consider testing the null hypothesis H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

(i) Give an example of a statistical test for H_0 for which the probability of Type II error is 100%. **(1p)**

(ii) Give an example of a statistical test for H_0 for which the probability of Type I error is 50%. **(1p)**

Hint: In each case, you do not need to care about the other error type, and the test does not necessarily need to be a conventional statistical test (although it can be). You need to describe a procedure that makes accept/reject decisions and has the required error rate.

a) The probability of a type 1 error (rejecting a true null hypothesis) can be minimized by **picking a smaller level of significance α before doing a test** (requiring a smaller p -value for rejecting H0) => Control the probability of type I error by modifying the significance level alpha
b) Type II error rate is more difficult to control as it is usually a function of the possible distributions under the alternative hypothesis H1
c) When consider adjusting the significance level (with Bonferroni correction) in hypothesis: in multiple testing problems such as in ANOVA or two sample rank test (Wilcoxon rank sum test) when the null hypothesis is rejected and we need to find out which pair is statistically significant. Why we need to do so: conducting all possible comparisons has the side effect that the probability of type I error is inflated greatly above its set level. Thus, to be absolutely sure that the probability of making at least one type I error during the C tests is at most some α, the individual comparisons must be done on significance level β = α/C.
d) An example of a statistical test for H0 for which the probability of type II error is 100% => This statistical test always accept the false null hypothesis => The statistical model always assume that the sample is normally distributed even if the sample is not normally distributed. Example: using t-test on very few datapoints which doesnt assume normality, meanwhile the t-test assumes the data to be normally distributed
An example of a statistical test for H0 for which the probability of type I error is 50% => This statistical test has a half probability of rejecting the true null hypothesis => The statistical model with 50% prob assume that the sample is not normally distributed even if the sample is normally distributed.

April 18, 2019

1. Answer either TRUE or FALSE (**1p** per item for correct answer, maximum amount of points obtainable is 6).
   a. In two-sample proportion test the sample sizes of the two groups need not be the same.
   b. In the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the error term is usually assumed to have expected value one, $\mathrm{E}(\varepsilon_i) = 1$.
   c. The aim of descriptive statistics is to draw conclusions about a population based on a sample.
   d. The two-sample rank test (Wilcoxon rank-sum test) makes the assumption that the medians of the distributions of the two samples are the same.
   e. Bartlett's test is a normality test (that is, used to test whether a sample comes from a normal distribution).
   f. Median is a measure of scatter.
   g. LASSO can be used for variable selection.
   h. In bootstrap, the number of observations in each of the bootstrap samples is the same as the number of observations in the original sample.

(a) True
(b) False, it is assumed to be 0
c) False, aim of descriptive statistics is to provide a brief summary of the sample. Drawing conclusions about the population based on a sample is the purpose of inferential statistics
(d) False. That is its null hypothesis. Its assumption is that x is i.i.d from a continuous and symmetric distribution
(e) False. It tests if groups have similar variance if they all follow normal distribution

Bartlett's test for equality of variances

Bartlett's test, assumptions

Let $x_{1j}, x_{2j}, \ldots, x_{n_j j}$ be i.i.d. observed values of a $\mathcal{N}(\mu_j, \sigma_j^2)$-distributed random variable x_j, $j = 1, \ldots, k$. Assume that the k samples are independent.

Bartlett's test, hypotheses

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

$H_1 : \sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

(f) False. Median absolute deviation is a measure of scatter. Median is a measure of location
(g) True. Ridge regression however cannot be used for variable selection
(h) True

2. a. Assume you are testing the following pair of hypotheses with some method of normality testing,

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution
H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution

Describe what it means to conduct Type I and Type II errors *in this context* (do not give the general definitions of Type I and II errors but instead state what they mean for this specific pair of hypotheses). **(2p)**

b. Draw an example of a quantile-quantile (Q-Q) plot where:
   i. the sample clearly comes from a normal distribution, **(1p)**
   ii. the sample clearly does not come from a normal distribution. **(1p)**

c. Name two different ways besides Q-Q plot for checking/testing the normality of a sample. **(1p)**

d. A researcher wants to model her data with *Model X* that makes a normality assumption. For this, she tests her data for normality and gets a p-value of 0.055 (for the hypotheses given in part a). Based on the p-value, she decides to use *Model X*. Can the researcher fully trust the results of the model? Explain why or why not. **(1p)**

(Solved in next year model exam)

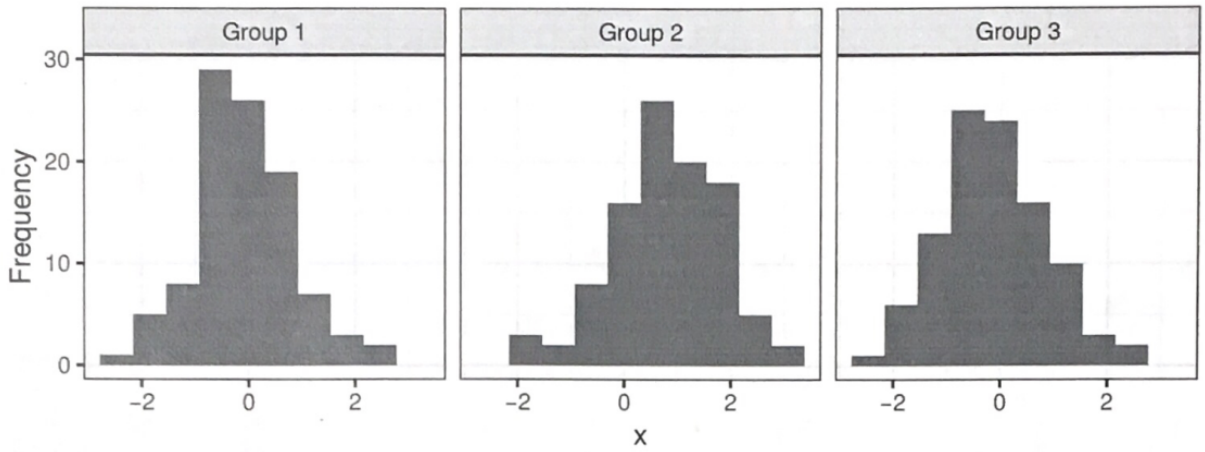
3. Consider analysis of variance (ANOVA) on a sample of three groups with 50 observations in each of them (assume that the groups are independent and that the observations are i.i.d. within each group). On the next page are shown the histograms of the groups, ANOVA summary and the results of Bartlett's test.

a. State the null hypothesis and the alternative hypothesis of ANOVA for this three-group case. **(1p)**

b. What would you conclude based on the ANOVA results? **(1p)**

c. Describe how one can check whether the assumptions of ANOVA are satisfied. Are they satisfied in the current example? **(2p)**

d. The next step in the analysis would be to conduct pair-wise testing between the groups. Bonferroni correction is often used in this context. Why is this? Describe also how the Bonferroni correction is applied. **(2p)**




```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## group         2  57.83  28.913   31.61 3.57e-13 ***
## Residuals   297 271.63   0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## ################################################################

##
##  Bartlett test of homogeneity of variances
##
## data:  x by group
## Bartlett's K-squared = 0.85135, df = 2, p-value = 0.6533
```

a) The null hypothesis of ANOVA for three-group cases: $\mu_1 = \mu_2 = \mu_3$

The alternative hypothesis of ANOVA for three-group cases: $\mu_i \neq \mu_j$ for some $i \neq j$

b) We have F = 31.61. The expected value of F test under H0 is F = (n - k)/(n - k - 2) = (50 x 3 - 3)/(50 x 3 - 3 - 2) = 1.013. Since F is very large from the expected value under the null hypothesis and the p value Pr(>F) is extremely small => The ANOVA test states that there is evidence that the expected values of at least two groups differ

c)

ANOVA has 2 key assumptions: The groups are normally distributed and Groups have equal variances. The first one can be tested by using Q-Q plot or histogram , the second one can be tested by Bartlett's test.

From the histograms, the group-wise normality seems plausible. The p-value of Bartlett's test is greater than 0.05, therefore showing evidence that the variances of the groups do not differ.

=> The normality assumptions of ANOVA are satisfied in this case

d) The analysis could be continued with pair-wise testing using the two-sample rank test to find out which pairs of groups have differing medians accompanied with the Bonferroni correction.

Multiple testing problem

If the null hypothesis is rejected based on the F-test, then we know that at least two of the groups differ (but we do not know which ones!).

The next step in the analysis is usually to find out the groups with statistically significant differences in expected values.

A simple way to do this is to analyze the groups in pairs of two with t-test.

There are $C = \frac{k(k-1)}{2}$ pairs in total to compare and conducting all possible comparisons has the side effect that the probability of type I error is inflated greatly above its set level.

This is called the *multiple testing problem*.

Bonferroni correction

Let β be the significance level of the C pair-wise comparisons, i.e, the (upper bound for the) probability that H_0 is incorrectly rejected in a single comparison, i.e., the probability of type I error in a single comparison.

Let γ be the probability that H_0 is incorrectly rejected in at least one test, when the test is repeated C times, i.e., the probability of making at least type I error during the C tests.

Probability theory shows that,

- if the tests are independent (which they most likely are not), then $\gamma = 1 - (1 - \beta)^C$.
- in the general case, we have the bound $\gamma \leq C\beta$.

Thus, to be absolutely sure that the probability of making at least one type I error during the C tests is at most some α, the individual comparisons must be done on significance level $\beta = \frac{\alpha}{C}$.

4. a. Explain the difference between errors and residuals in a linear regression model. **(1p)**
   b. Give two uses for the residuals of a linear regression model. **(2p)**
   c. What does multicollinearity mean? **(1p)**
   d. Consider a drug experiment where
      - the continuous response y_i is the change in the amount of a specific antigen in the ith patient's blood one day after receiving the drug (higher is better),
      - the binary predictor x_{i1} describes which drug the ith patient received ($x_{i1} = 0$ for placebo, $x_{i1} = 1$ for the new experimental substance),
      - the continuous predictor x_{i2} describes the amount of the drug (placebo or the new experimental substance) the ith patient received.

      To study whether the new experimental substance is more efficient than placebo in increasing the amount of the antigen in blood, we fit the linear regression model

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}.$$

      Note: "placebo" refers to a sugar pill or something similar which should have no effect on the patient.

      i. Which null hypothesis (concerning the model coefficients $\beta_0, \beta_1, \beta_2, \beta_{12}$) should we test to determine whether the new experimental substance and the placebo are equally effective in increasing the amount of the antigen in blood? **(1p)**

      ii. It seems reasonable to assume that for those patients who received placebo, the amount of placebo received has no effect on the outcome. State this observation in terms of the model coefficients $\beta_0, \beta_1, \beta_2, \beta_{12}$. **(1p)**

a) **Error** of the data set is the differences between the observed values and the true / unobserved values. **Residual** is calculated after running the regression model and is the difference between the observed values and the estimated values. Moreover, Residual can be considered as an estimate of error.
• The error of an observed value is the deviation of the observed value from the true value of a quantity of interest (for example, a population mean).
• The residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean). The distinction is most important in regression analysis, where the concepts are sometimes called the regression errors and regression residuals.

b) Residual shows how well the regression model explain the observed variables of the response variable. It measure the difference between observed variable and fitted variable.
- One use is to help us to determine if we have a data set that has an overall linear trend, or if we should consider a different model. The reason for this is that residuals plot help to amplify any nonlinear pattern in our data
- Another reason to consider residuals is to check that the conditions for inference for linear regression are met. After verification of a linear trend (by checking the residuals), we also check the distribution of the residuals. In order to be able to perform regression inference, we want the residuals about our regression line to be approximately normally distributed. A histogram or stemplot of the residuals will help to verify that this condition has been met.

c) In statistics, multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

d) Drug experiment

(i) Model: $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$

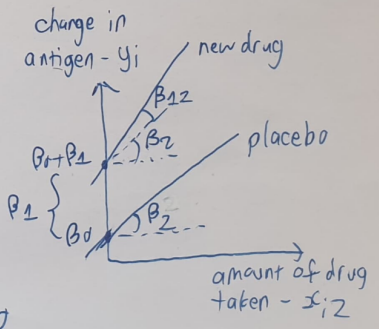
Placebo: $x_{i1} = 0 \Rightarrow E(y_i) = \beta_0 + \beta_2 x_{i2}$

tested drug: $x_{i1} = 1 \Rightarrow E(y_i) = \beta_0 + \beta_1 + \beta_2 x_{i2} + \beta_{12} x_{i2}$

$= (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) x_{i2}$

$\Rightarrow$ If the new tested drug and the placebo have the same effectivness $\Rightarrow$ the intercept and slope of their regression line should be the same

$\Rightarrow \begin{cases} \beta_0 = \beta_0 + \beta_1 \\ \beta_2 = \beta_2 + \beta_{12} \end{cases} \Rightarrow \begin{cases} \beta_1 = 0 \\ \beta_{12} = 0 \end{cases} \Rightarrow$ H_0 we need to test is $\begin{cases} \beta_1 = 0 \\ \beta_{12} = 0 \end{cases}$



(ii) Placebo has no effect no matter how much it is taken

$\Rightarrow$ The change in antigen y_i does not depend on amount of drug taken $x_{i2} \Rightarrow$ slope of the model is 0

$\Rightarrow \beta_2 = 0$, Assume that the antigen does not naturally increase after one day $\Rightarrow \beta_0 = 0$

There are no constraints on β_1 and β_{12}. When $\beta_2 = 0, \beta_0 = 0$

$\Rightarrow$ model of tested drug: $\beta_1 + \beta_{12} x_{i2} = E(y_i) \Rightarrow \beta_1$ & β_{12} will be intercept and slope for the tested drug

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
13.4.2022
J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Allowed equipment: writing tools, calculator (symbolic and graphic OK), at most A4-size cheat sheet written on one side.

P1 In each question, answer TRUE or FALSE. In this problem you do not need to write reasons. 0.5 points per item for correct answer. (6p)

(a) Descriptive statistics aims to draw conclusions about a population based on a sample. F

(b) Sample mean is the 0.5-quantile of a sample. F

(c) If a distribution is symmetric around zero, it is called the normal distribution. F

(d) The chi-squared (χ^2) distribution is symmetric around zero. F

(e) Positive skewness indicates that the distribution has a long right tail. T

(f) In linear regression, correlation coefficient is the slope of the regression line. F

(g) Confidence level indicates how much confidence you have in the model assumptions. F

(h) In a simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the error term ϵ_i is usually assumed to have zero expected value. T

(i) Bootstrapping is a method for using linear regression with multiple predictor variables. F

(j) Signed rank tests make stricter assumptions than sign tests. T

(k) In hypothesis testing, the null hypothesis is rejected when the p-value is greater than the significance level. F $p > \alpha \Rightarrow H_0$ ✓

(l) In multiple testing, Bonferroni correction increases the probability of Type II errors. F

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give their general definitions but explain what they mean here specifically). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name and briefly explain two ways besides Q-Q plot for checking/testing the normality of a sample. **(2p)**

a) b) Easy

c) Two ways are Shapiro Wilk test and Bowman Shenton test

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
13.4.2022
J Kohonen

P3 (a) Draw a scatter plot of two variables that have:

(i) perfect linear dependence **(1p)**

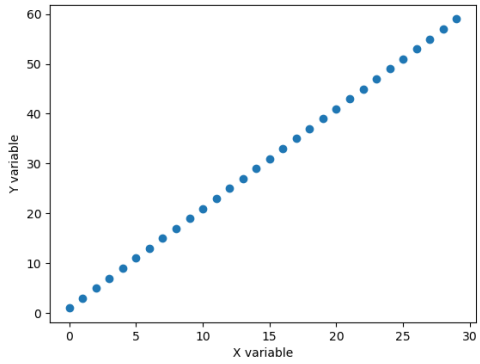
(ii) perfect monotonic dependence but not perfect linear dependence **(1p)**

(b) Is it possible for two variables to have perfect linear dependence but not perfect monotonic dependence? Explain why or why not. **(2p)**

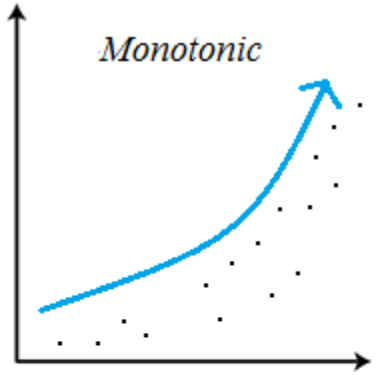
(c) Explain Spearman's rank correlation coefficient. When is it used and how? **(2p)**

a)

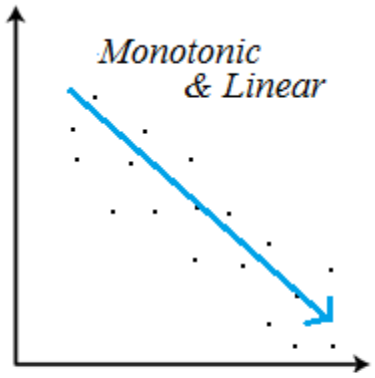
Perfect linear dependence



Perfect monotonic dependence but not perfect linear dependence



b) Not possible because linear dependence already implies monotonic dependence.



c)

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

Spearman's rank correlation coefficient

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).
- Let $R(x_i)$ denote the rank of the observation x_i in the sample $x_1, x_2, \ldots, x_n$ and let $R(y_i)$ denote the rank of the observation y_i in the sample $y_1, y_2, \ldots, y_n$.
- Then Spearman's rank correlation coefficient $\rho_S(x, y)$ is the Pearson's correlation coefficient calculated from the ranks.

- Spearman's rank correlation coefficient measures the monotonic dependence between two random variables. The coefficient is always in the interval $[-1, 1]$ and (in case of no repeating data values) attains the absolute value 1 if and only if $y = g(x)$ for some monotonic function g.
- If the variables x and y are independent, then the Spearman correlation $\rho_S(x, y) = 0$ (using the same counterexample as with Pearson correlation, we see that the contrary does not again hold).

P4 Consider a dataset of $n = 7$ observations with x values

$$3.0, 3.5, 4.0, 5.0, 5.5, 6.0, 9.0$$

and y values

$$0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 2.0.$$

Let K be the triangular kernel function

$$K(u) = \begin{cases} 1 - |u| & \text{if } |u| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Calculate the Nadaraya-Watson regression function at points $x = 4.0, 4.1, 4.5, 4.9, 5.0$ and 7.5 using K as the kernel. Give at least three decimals. If at some points it cannot be calculated, explain why. **(2p)**

(b) What happens with this dataset if we use the kernel function $K_{0.01}(u) = K(u/0.01)$? Where exactly can the regression be calculated and what are its values there? **(1p)**

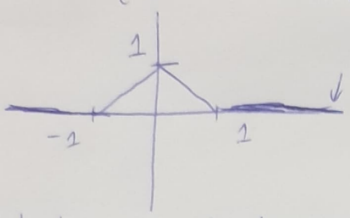
(c) Calculate the regression function at $x = 4.0$ and at $x = 7.5$ using the kernel function $K_{100}(u) = K(u/100)$. Give at least three decimals. **(1p)**

(d) Explain in general terms (not just for this dataset) how bandwidth affects Nadaraya-Watson kernel regression. Consider small, intermediate and large values. **(1p)**

(e) Explain how cross-validation can be used to select bandwidth in kernel regression. **(1p)**

a)

$$K(u) = \begin{cases} 1 - |u| & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$



$x = [3.0, 3.5, 4.0, 5.0, 5.5, 6.0, 9.0]$

$y = [0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 2.0]$

a) Calculating the Nadaraya-Watson regression functions at point 4.0, 4.1, 4.5, 4.9 5.0 and 7.5 using K as the kernel

- Nadaraya-Watson regression: At any point x, define regression function value as a weighted average of data points: $g(x) = \sum_{i=1}^{n} w_i y_i$ where the weights are:

$$w_i = \frac{K(x - x_i)}{\sum_{j=1}^{n} K(x - x_j)}, \text{ where } \sum_{i=1}^{n} w_i = 1$$

Now, for $x = 4.0$

We have: $\sum_{j=1}^{n} K(x - x_j) = K(4-3) + K(4-3.5) + K(4-4) \ldots$

$= K(1) + K(0.5) + K(0) + \ldots$

$= 0 + 0.5 + 1 + 0 + 0 + 0 + 0 = 1.5$

$\Rightarrow w_1 = \frac{K(4-3)}{1.5} = \frac{0}{1.5} = 0$

$\Rightarrow w_i = [0, 1/3, 2/3, 0, 0, 0, 0]$

$\Rightarrow g(4) = w_1 y_1 + w_2 y_2 + \ldots = [0\ 1/3\ 2/3\ 0\ 0\ 0\ 0] \times [0\ 0\ 0\ 1\ 1\ 1\ 2] = 0$

- For $x = 4.1$: $\sum_{j=1}^{n} K(x - x_j) = 0 + 0.6 + 0.1 + 0.9 + 0 + 0 + 0 = 1.6$

$\Rightarrow w_i = [0\ 3/8\ 1/16\ 9/16\ 0\ 0\ 0]$

$y_i = [0\ 0\ 0\ 1\ 1\ 1\ 2]$ $\Rightarrow g(4.1) = 9/16$

- For $x = 4.5$: $\sum_{j=1}^{n} K(x - x_j) = [0\ 0\ 0.5\ 0.5\ 0\ 0\ 0] = 1$

$\Rightarrow w_i = [0\ 0\ 1/2\ 1/2\ 0\ 0\ 0]$

$y_i = [0\ 0\ 0\ 1\ 1\ 1\ 2]$ $\Rightarrow g(4.5) = 1/2$

- For $x = 4.9$: $\sum_{j=1}^{n} K(x - x_j) = [0\ 0\ 0.9\ 0.1\ 0.6\ 0\ 0] = 1.6$

$\Rightarrow w_i = [0\ 0\ 9/16\ 1/16\ 3/8\ 0\ 0]$

$y_i = [0\ 0\ 0\ 1\ 1\ 1\ 2]$ $\Rightarrow g(4.9) = 7/16$

- For $x = 5.0$ $\sum_{j=1}^{n} K(x - x_j) = [0\ 0\ 0\ 0\ 0.5\ 0\ 0] = 0.5$

$\Rightarrow w_i = [0\ 0\ 0\ 0\ 1\ 0\ 0]$

$y_i = [0\ 0\ 0\ 1\ 1\ 1\ 2]$ $\Rightarrow g(5.0) = 1$

- For $x = 7.5$: $\sum_{j=1}^{n} K(x - x_j) = [0\ 0\ 0\ 0\ 0\ 0\ 0] = 0$

$\Rightarrow$ Cannot be calculated because denominator is 0 in $w_i = \frac{K(x - x_i)}{\sum_{j=1}^{n} K(x - x_j)}$

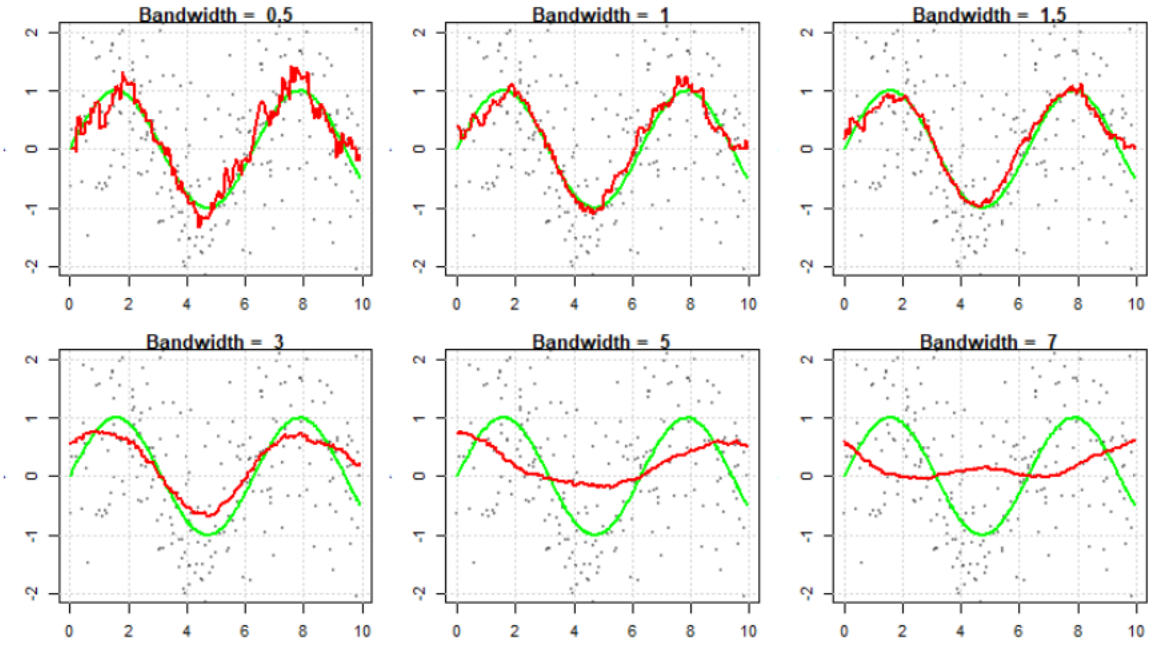
b)

The choice of bandwidth λ is critical to the estimator's performance and far more important than the choice of kernel. If the smoothing parameter is too small, the estimator will be too rough; but if it is too large, we risk smoothing out important function features. In other words, choosing λ involves a significant bias-variance trade-off.

$\lambda \uparrow \quad \Rightarrow$ smooth curve, low variance, high bias

$\lambda \downarrow \quad \Rightarrow$ rough curve, high variance, low bias

As the bandwidth is 0.01 which is very small, we will have very rough curve with very high variance



Kernel regression fits for various values of λ.

Bandwidth in kernel regression is called **the smoothing parameter because it controls variance and bias in the output**.

The regression can be calculated in the interval of [-h, h] or [-0.01, 0.01]. Its values there is
1 - abs(u/0.01)

Choice of kernel function

The kernel function is typically defined in two steps:

1. Choose a shape, such as a triangular function, parabola, or the density function of standard normal distribution
2. Choose a bandwidth that scales the shape to desired width = how far datapoints are used in the averaging

Example: parabolic (Epanechnikov) kernel

$$K_1(u) = \frac{3}{4}(1 - u^2)$$

for $-1 \leq u \leq 1$, and zero outside that interval.
Then scaled to bandwidth h with

$$K_h(u) = K_1(u/h).$$

This is positive for $-h \leq u \leq h$.
See https://en.wikipedia.org/wiki/Kernel_(statistics) for many other kernel shapes.

Local linear regression

Instead of taking the average of the nearby points, we can also fit a straight line to them. This is called local linear regression.

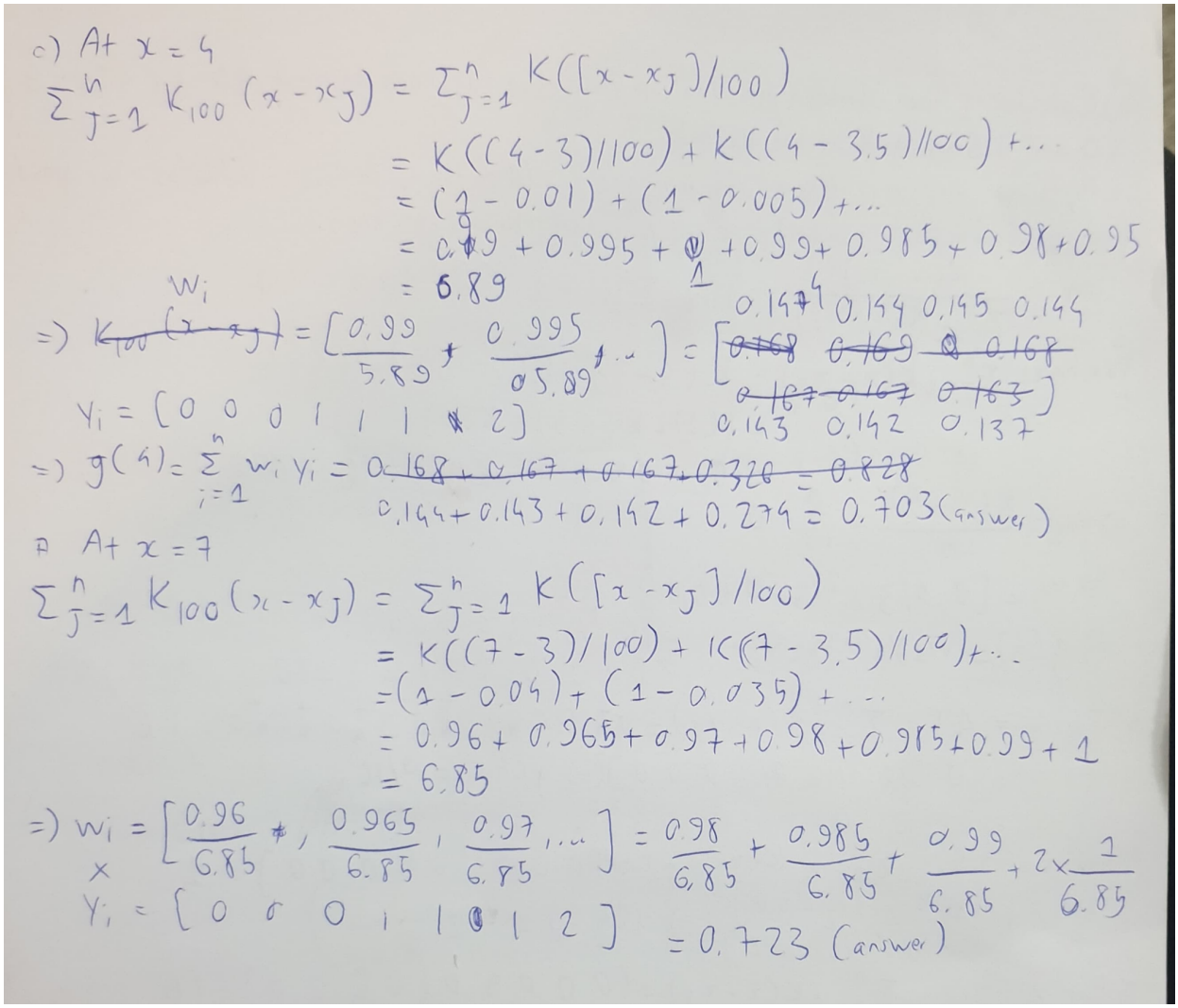
In other words, we do a linear regression, but only on the data points x_i that are near x, and weighted by a kernel function. Then define $g(x)$ to be the value of that regression line at x.

Note that for each value x where we are evaluating the regression function, we look at *different* "nearby" datapoints or use different weights, so the regression function $g(x)$ that we obtain need not be "linear" at all.

Nadaraya-Watson (local constant) and local linear regression usually produce similar results, except at edges of the data. (Consider what happens in time series prediction.)

Just like in "global" regression, in local regression we can also use higher degree polynomials (e.g. parabolic).

https://statisticelle.com/ml-theory-kernel-regression/
c)



* typo: divided by 6.89 in x = 4, not 5.89

d)

Choice of bandwidth

Large bandwidth = averaging many datapoints = very smooth regression function that only shows "large scale" features of the data. Also efficiently smoothes small errors away.

Small bandwidth = averaging few datapoints = very wild regression function that follows the data very closely. But also retains its errors.

Many methods exist for choosing the "best" bandwidth (see literature), but for exploratory analysis you could just experiment with different values. There are also "adaptive" methods which use smaller bandwidth if there are many data points nearby.

e)

Cross-Validation Methods

Selecting the amount of smoothing using subjective methods requires time and effort. Automatic selection of λ can be done via cross-validation. The cross-validation criterion is

$$CV(\lambda) = \frac{1}{n} \sum_{n} \left(y_j - \hat{m}_{\lambda}^{(-j)} x_j \right)^2$$

where $(-j)$ indicates that point j is left out of the fit. The basic idea is to leave out observation j and estimate $m(\cdot)$ based on the other $n-1$ observations. λ is chosen to minimize this criterion.

True cross-validation is computationally expensive, so an approximation known as generalized cross-validation (GCV) is often used. GCV approximates CV and involves only one non-parametric fit for each λ value (compared to CV which requires n fits at each λ).

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
1.6.2022
J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Allowed equipment: writing tools, calculator (symbolic and graphic OK), at most A4-size cheat sheet written on one side.

P1 In each question, answer true or false. 1 point per correct answer.

(a) Sample median is the 0.5-quantile of a sample. T

(b) If a distribution is symmetric around zero, it is called the normal distribution. F

(c) Positive skewness indicates that the expectation of the distribution is above zero. F

(d) In linear regression, correlation coefficient is the slope of the regression line. F

(e) Confidence level indicates how much confidence you have in the model assumptions. F

(f) Bonferroni correction is used for removing outliers from a sample. F

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give their general definitions but explain what they mean here specifically). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name and briefly explain two ways besides Q-Q plot for checking/testing the normality of a sample. **(2p)**

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
1.6.2022
J Kohonen

P3 Consider multiple linear regression on a sample of $n = 100$ observations of a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. Below are shown the linear regression model summary, variance inflation factors and the diagnostics plot (next page) for the model fit.

(a) Describe how one can check whether the assumptions of multiple linear regression are satisfied. Are they satisfied in the current example? **(3p)**

(b) Based on the variance inflation factors, can the estimated coefficients be trusted? Why or why not? Explain what it would mean if these factors are high or low. **(1p)**

(c) Give an interpretation for the estimated coefficient $\hat{\beta}_4 \approx 0.21811$. **(1p)**

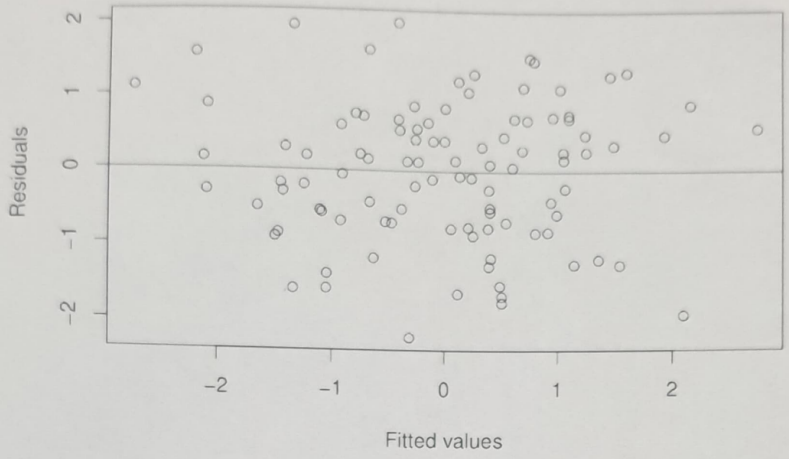
(d) What does the fitted model predict for the response variable if $x_{i1} = x_{i2} = x_{i3} = 0$, $x_{i4} = 100$, and $x_{i5} = 0$? Give a numerical answer and explain why or why not this prediction can be trusted. **(1p)**

```
## Call:
## lm(formula = y ~ ., data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2158 -0.6744  0.1267  0.6918  1.9987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08357    0.09932  -0.841   0.4022
## x1          -0.59245    0.09081  -6.524 3.42e-09 ***
## x2          -0.07999    0.12061  -0.663   0.5088
## x3          -0.55658    0.21660  -2.570   0.0118 *
## x4           0.21811    0.20657   1.056   0.2937
## x5           0.72269    0.09904   7.297 9.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9499 on 94 degrees of freedom
## Multiple R-squared:  0.5637, Adjusted R-squared:  0.5405
## F-statistic: 24.29 on 5 and 94 DF,  p-value: 1.285e-15

## Variance inflation factors
##       x1       x2       x3       x4       x5
## 1.014837 1.024145 5.757164 5.738664 1.056317
```

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
1.6.2022
J Kohonen



P4 Consider a dataset of $n = 7$ observations with x values

$$3.0, 3.5, 4.0, 5.0, 5.5, 6.0, 9.0$$

and y values

$$0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 2.0.$$

Let K be the **rectangular** kernel function

$$K(u) = \begin{cases} 1 & \text{if } |u| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

0

(a) Calculate the Nadaraya-Watson regression function at points $x = 3.5, 4.2, 7.5$ and 9.2 using K as the kernel. Give at least three decimals. If at some points it cannot be calculated, explain why. **(2p)** 1/3 undefined 2

(b) What happens with this dataset if we use the kernel function $K_{0.01}(u) = K(u/0.01)$? Where exactly can the regression be calculated and what are its values there? **(1p)**

(c) Calculate the regression function at $x = 4.0$ and at $x = 7.5$ using the kernel function $K_{100}(u) = K(u/100)$. Give at least three decimals. **(1p)** 5/7 0.714

(d) Explain in general terms (not just for this dataset) how bandwidth affects Nadaraya-Watson kernel regression. Consider both small and large values. **(1p)**

(e) Explain how cross-validation can be used to select bandwidth in kernel regression. **(1p)**