THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

This exam paper must not be removed from the venue

Venue _____

Seat Number _____

Student Number |__|__|__|__|__|__|__|__|

Family Name _____

First Name _____

School of Mathematics & Physics

EXAMINATION

Semester Two Final Examinations, 2021

STAT2203 Probability Models and Data Analysis

This paper is for St Lucia Campus students.

Examination Duration: 120 minutes

Reading Time: 10 minutes

Exam Conditions:

This is a Closed Book examination - specified written materials permitted

Casio FX82 series or a calculator on the UQ approved list

During planning time - students are encouraged to review and plan responses to the exam questions

This examination paper will be released to the Library

Materials Permitted In The Exam Venue:

(No electronic aids are permitted e.g. laptops, phones)

One A4 sheet of handwritten notes double sided is permitted

Materials To Be Supplied To Students:

None

Instructions To Students:

Additional exam materials (eg. answer booklets, rough paper) will be provided upon request.

There are 60 marks available on this exam from 6 questions.

Write your answers in the spaces provided on pages 2 - 16 of this examination paper. Show your working and state conclusions where appropriate. Some probabilities and quantiles from R are provided on Page 17. Some useful formulas are provided on Page 18 - 19.

For Examiner Use Only

Question	Mark

Total _____

1. [8 marks] The ex-Gaussian distribution is used in experimental psychology to model response times. Suppose U and V are independent random variables such that $U \sim \text{Exp}(1)$ and $V \sim \mathcal{N}(0,1)$. Then the random variable $X = 20U + 5V$ has an ex-Gaussian distribution.

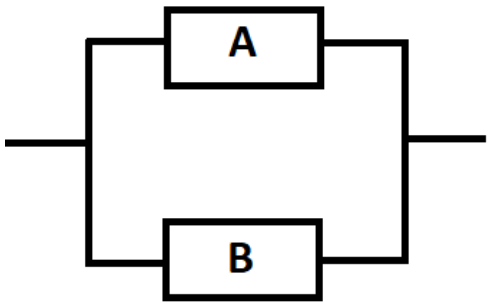
   (a) Compute $\mathbb{E}X$. [2 marks]

   (b) Compute $\text{Var}(X)$. [2 marks]

   (c) Compute $\text{Cov}(X, V)$. [2 marks]

   (d) Compute the moment generating function of X. [2 marks]

2. [12 marks] Consider a system comprising two components (A and B) connected in parallel.



The system is working if there is a path from left to right through working components. The cumulative distribution function of the time to failure for each component is

$$F(t) = \begin{cases} 1 - t^{-3}, & t \geqslant 1 \\ 0, & \text{else} \end{cases}$$

(a) Determine the quantile function of F. [3 marks]

(b) Describe how random variables with cumulative distribution function F can be simulated given a means of generating random variables from a $\mathcal{U}[0, 1]$ distribution.

[1 mark]

(c) What is the probability that component A will not fail in its first 5 years of operation, given it has been operating for at least 2 years. [3 marks]

(d) Assuming the time to failure for components A and B are independent, determine the probability density function for the time to failure of the system. [5 marks]

3. [10 marks] A pair of random variables (X, Y) has a joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \exp(-x(1+y)^2), & x > 0, y > 0 \\ 0, & \text{else} \end{cases}$$

(a) Determine the marginal probability density function of Y. [2 marks]

(b) What is the conditional distribution of X, given $Y = 2$. [2 marks]

(c) Are X and Y independent random variables? Justify your answer. [2 marks]

(d) Compute $\mathbb{E}[XY]$.

[4 marks]

4. [13 marks] One hundred and eighteen fourth-grade children from four American public schools each completed a survey about their video game playing habits. Students were asked about their preferred genre (Action, Adventure, Simulation), time spent playing video games, and the strategies they used to improve at the games they play most often.

   (a) The 66 students that preferred Action video games spent an average of 3.98 hours per week playing video games, with a sample standard deviation of 2.90 hours. The 22 students that preferred Simulation video games spent an average of 2.57 hours per week playing video games, with a sample standard deviation of 1.93 hours.

   Is there any evidence of a difference in the mean time spent playing video games between students that prefer Action video games and students that prefer Simulation video games? State the null and alternative hypotheses, and use an appropriate test statistic to determine the p-value. What do you conclude?

   [4 marks]

(b) Overall, the 118 students surveyed averaged 3.55 hours per week playing video games, with a sample standard deviation of 2.99 hours. Construct a 95% confidence interval for the mean hours per week spent playing video games of all students. [3 marks]

(c) Out of 118 students surveyed, 66 students preferred the action genre of video games. Construct a 95% confidence interval for the proportion of all students who prefer the action genre. [3 marks]

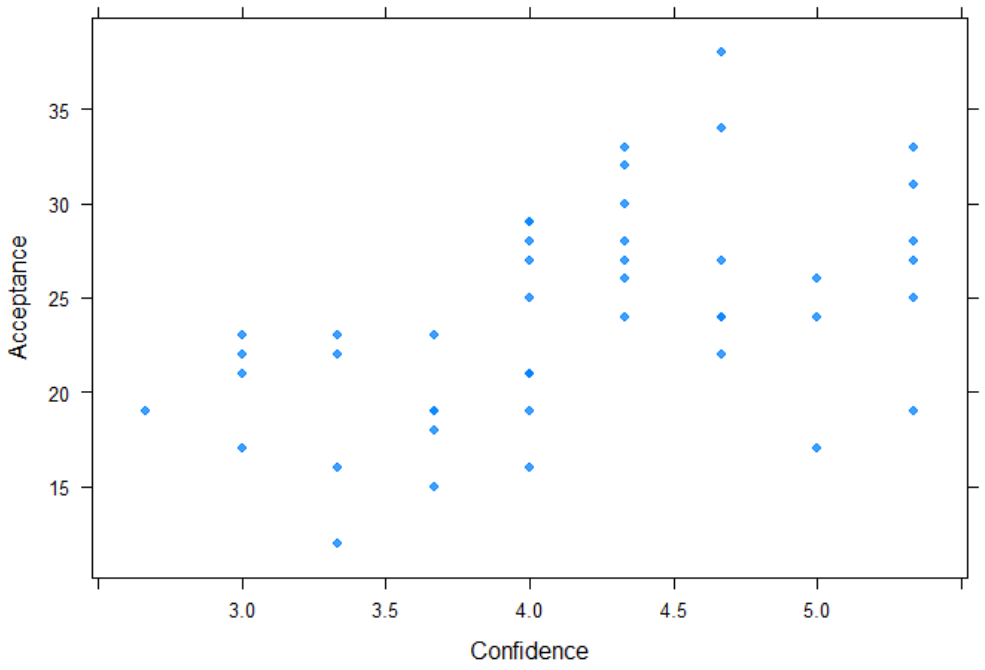
(d) One strategy students used to improve at the video games was "I play it over and over again", referred to as repetition. The table below shows the students preferred video game genre and whether or not they used the repetition strategy.

Repetition	Genre			Total
	Action	Adventure	Simulation	
Yes	46	12	12	70
No	20	18	10	48
Total	66	30	22	118

i. Assuming the students' preferred genre of video games and the strategy used to improve are independent, what would be the expected count for students who prefer Action video games and use the repetition strategy? [1 mark]

ii. We want to know if there is any evidence of dependence between the students preferred video game genre and the strategy they use to improve. The appropriate analysis in `R` reported a test statistic of 7.7934. Determine the p-value for the test. What do you conclude? [2 marks]

5. [10 marks] Care workers were surveyed to understand factors associated with the acceptance of service robots by care workers to assist with providing home care to customers. The care worker's *Acceptance* of service robots was scored on a 45 point scale using a questionnaire with a higher score indicating greater acceptance of service robots. *Confidence* of the care worker in their ability to learn to use the service robots was also scored (range 1 - 6) using a questionnaire with a higher score indicating greater confidence. The *Age* of the care worker was also recorded.



A multiple regression model for *Acceptance* using *Confidence* and *Age* was fitted. The edited output from `R` after fitting is given below.

```
lm(formula = Acceptance ~ Confidence + Age)

Coefficients:
            Estimate Std. Error
(Intercept)  4.29116    5.67495
Confidence   3.92836    1.01161
Age          0.06851    0.06750
---

Residual standard error: 4.92 on 43 degrees of freedom
Multiple R-squared:  0.264,Adjusted R-squared:  0.2298
F-statistic: 7.712 on 2 and 43 DF,  p-value: 0.001373
```

(a) The following figures were generated to check the assumptions underlying the multiple regression. State the assumptions of the multiple regression model and comment on their validity for this data with reference to the figures below.
[3 marks]

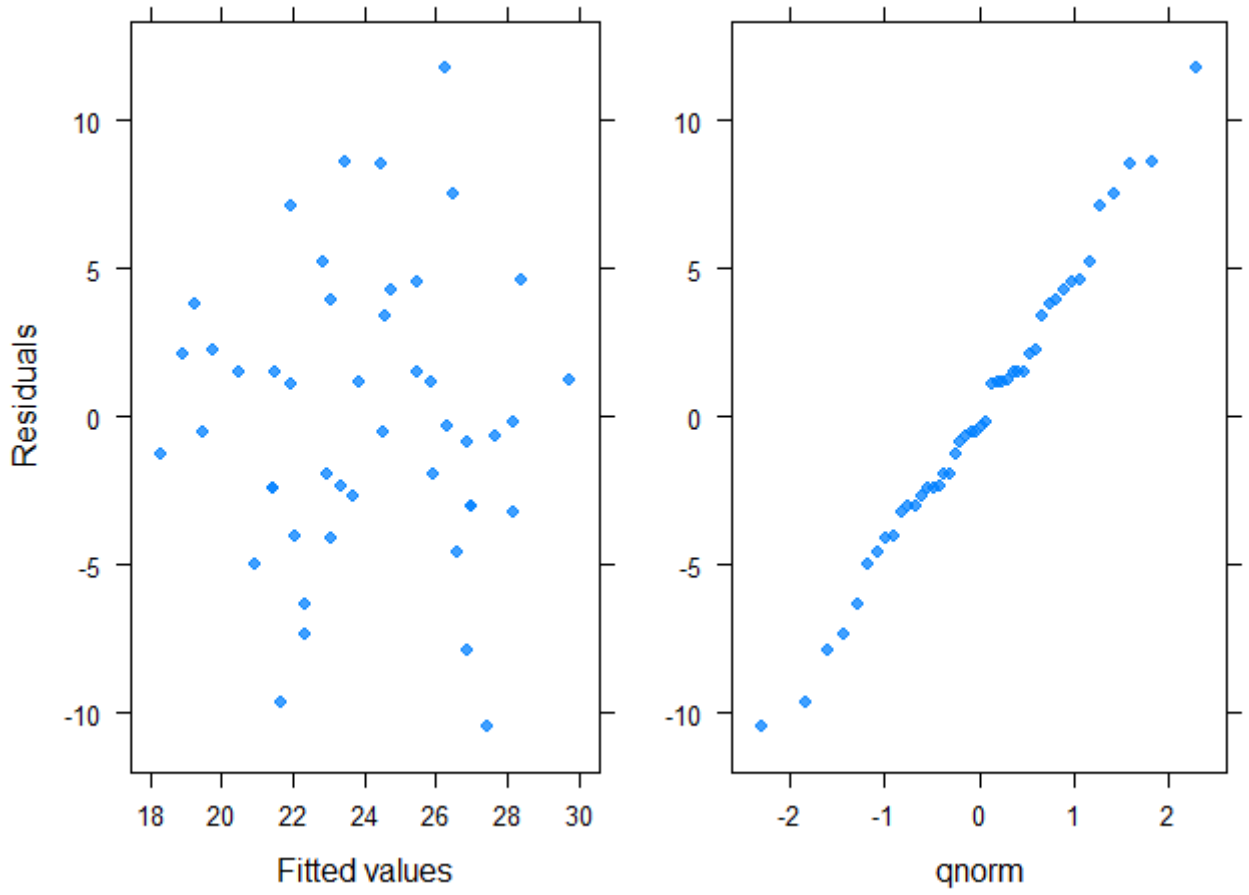


Figure 1: Left: Plot of residuals against fitted values from the linear regression. Right: Plot of residuals against quantiles of the standard normal distribution.

(additional space for answer to part (a))

(b) How many participants were in the study? [1 mark]

(c) What is the estimated *Acceptance* score for a care worker with a *Confidence* score of 3 and an *Age* of 40? [1 mark]

(d) Give a 95% confidence interval for the coefficient of *Confidence* in this multiple regression. [2 marks]

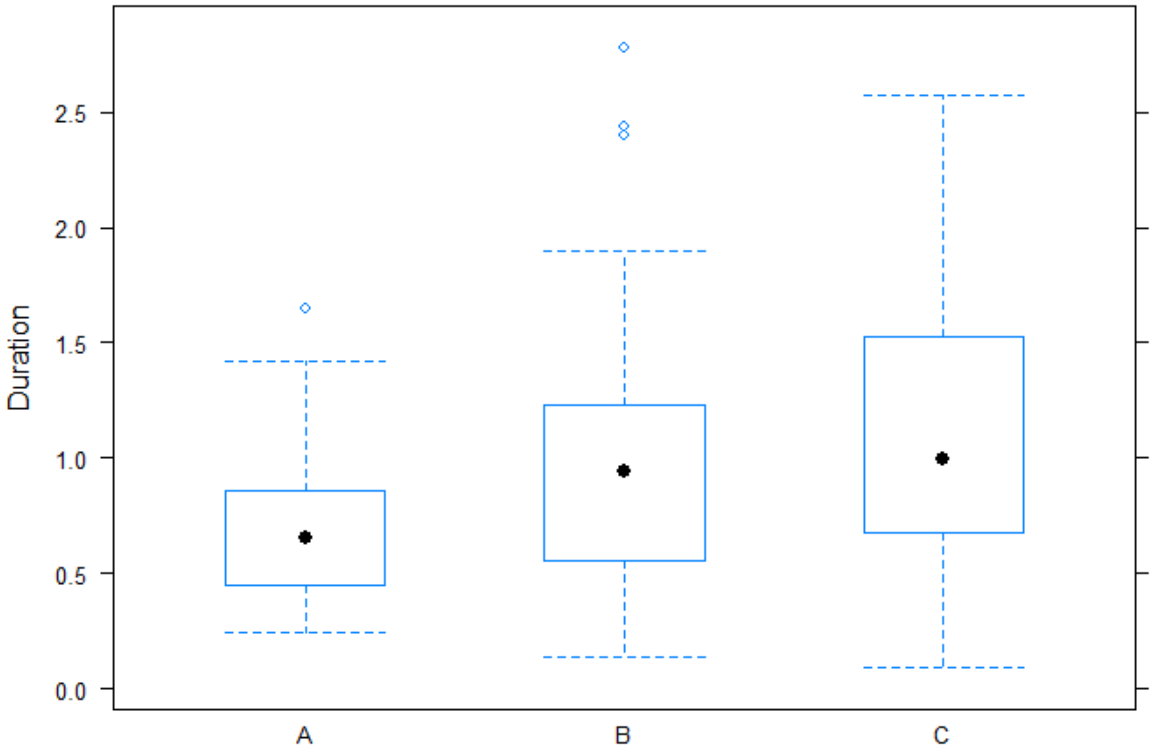
(e) Does the regression analysis provide evidence of an association between *Acceptance* score and *Age*, after taking into account *Confidence*?

[3 marks]

6. [7 marks] Researchers investigated the perceived time duration under three conditions: (A) reading a journal article, (B) watching a movie, and (C) occasional switching between reading a journal article and watching a movie. Seventy-five participants were recruited to the study and randomly allocated to one of the three groups (A, B, C). After 20 minutes the participants were asked to estimate the duration of time that had passed and the ratio of the perceived time passed to actual time passed was recorded. The data is summarised in the table below.

	n	$\bar{x}$	s
A	25	0.73	0.38
B	25	1.04	0.69
C	25	1.15	0.71

Below is a boxplot of the data.



(a) A one-way analysis-of-variance (ANOVA) is used to test whether there are systematic differences between groups in their mean perceived duration. Complete the following ANOVA table. [3 marks]

Source	DF	SS	MS	F
Groups		2.461		
		27.330		
Total				

(b) Determine the p-value for the hypothesis test. What do you conclude? [2 marks]

(c) How might the assumptions underlying the one-way analysis of variance be violated in this study? The boxplot on the previous page may help you address this question. [2 marks]

(additional space for answer to part (c))

END OF EXAMINATION

Probabilities and Quantiles

Normal distribution

```
> qnorm(0.95)
[1] 1.644854
```

```
> qnorm(0.975)
[1] 1.959964
```

```
> qnorm(0.995)
[1] 2.575829
```

t-distribution

```
> pt(1.863, df=21)
[1] 0.961745
> pt(1.863, df=22)
[1] 0.9620658
```

```
> pt(2.588, df=21)
[1] 0.9914196
> pt(2.588, df=22)
[1] 0.9916073
```

```
> pt(3.187, df=21)
[1] 0.9977824
> pt(3.187, df=22)
[1] 0.9978697
```

```
> qt(0.975, df=43)
[1] 2.016692
> qt(0.975, df=44)
[1] 2.015368
> qt(0.975, df=45)
[1] 2.014103
```

```
> pt(1.015, df= 43)
[1] 0.8421084
> pt(1.015, df= 44)
[1] 0.8421726
> pt(1.015, df= 45)
[1] 0.8422339
```

```
> pt(3.883, df= 43)
[1] 0.9998248
> pt(3.883, df= 44)
[1] 0.9998287
> pt(3.883, df= 45)
[1] 0.9998325
```

```
> qt(0.95, df=117)
[1] 1.657982
```

```
> qt(0.975, df=117)
[1] 1.980448
```

```
> qt(0.995, df=117)
[1] 2.618504
```

Chi-squared distribution

```
> pchisq(7.7934, df=1)
[1] 0.9947563
> pchisq(7.7934, df=2)
[1] 0.9796912
```

```
> pchisq(7.7934, df=3)
[1] 0.9495198
> pchisq(7.7934, df=5)
[1] 0.8320047
```

```
> pchisq(7.7934, df=6)
[1] 0.7463665
```

F distribution

```
> pf(2.131, 1, 73)
[1] 0.8513632
> pf(2.131, 2, 72)
[1] 0.8738536
> pf(2.131, 3, 71)
[1] 0.8960979
```

```
> pf(3.242, 1, 73)
[1] 0.9240963
> pf(3.242, 2, 72)
[1] 0.955141
> pf(3.242, 3, 71)
[1] 0.9729865
```

```
> pf(6.573, 1, 73)
[1] 0.9875882
> pf(6.573, 2, 72)
[1] 0.9976119
> pf(6.573, 3, 71)
[1] 0.9994478
```

Formula Sheet

Elementary probability

- **Sum rule:** For disjoint $A_1, A_2, \ldots$: $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$.
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- **Conditional probability:** $\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.
- **Law of total probability**: $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i)$, where $B_1, B_2, \ldots, B_n$ is a partition of Ω.
- **Bayes' Rule:** $\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\, \mathbb{P}(A|B_j)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\, \mathbb{P}(A|B_i)}$.
- **Independent events**: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\, \mathbb{P}(B)$.

Random variables

- **Cdf** of X: $F(x) = \mathbb{P}(X \leqslant x), \;\; x \in \mathbb{R}$.
- **Pmf** of X: (discrete r.v.) $f(x) = \mathbb{P}(X = x)$.
- **Pdf** of X: (continuous r.v.) $f(x) = F'(x)$.
- For a discrete r.v. X: $\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x)$.
- For a continuous r.v. X with pdf f: $\mathbb{P}(X \in B) = \int_B f(x)\, dx$.
- In particular (continuous), $F(x) = \int_{-\infty}^{x} f(u)\, du$.
- **Quantile function of F:** For $p \in (0, 1)$ $Q(p) = \inf\{x : F(x) \geqslant p\}$
- In particular (continuous), $F(Q(p)) = p$.
- **Important discrete distributions**:

Distr.	pmf	suppport
$\mathsf{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0, 1\}$
$\mathsf{Bin}(n, p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$\{0, 1, \ldots, n\}$
$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda}\, \frac{\lambda^x}{x!}$	$\{0, 1, \ldots\}$

- **Important continuous distributions**:

Distr.	pdf	$x \in$
$\mathcal{U}[a, b]$	$\frac{1}{b-a}$	$[a, b]$
$\mathsf{Exp}(\lambda)$	$\lambda\, \mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$
$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$

- **Expectation** (discr.): $\mathbb{E}X = \sum_x x\, \mathbb{P}(X = x)$.
- (of function) $\mathbb{E}\, g(X) = \sum_x g(x)\, \mathbb{P}(X = x)$.
- **Expectation** (cont.): $\mathbb{E}X = \int x\, f(x)\, dx$.
- (of function) $\mathbb{E}\, g(X) = \int g(x) f(x)\, dx$,
- $\mathbb{E}X$ **and** $\mathbf{Var}(X)$ **for discrete distributions**:

	$\mathbb{E}X$	$\mathrm{Var}(X)$
$\mathsf{Ber}(p)$	p	$p(1-p)$
$\mathsf{Bin}(n, p)$	np	$np(1-p)$
$\mathsf{Poi}(\lambda)$	λ	λ

- $\mathbb{E}X$ **and** $\mathbf{Var}(X)$ **for continuous distributions**:

	$\mathbb{E}X$	$\mathrm{Var}(X)$
$\mathcal{U}(a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathsf{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2

Multiple random variables

- **Joint distribution:** $\mathbb{P}((X, Y) \in B) = \iint_B f_{X,Y}(x, y)\, dx\, dy$ (cont.) $\mathbb{P}((X, Y) \in B) = \sum_B f_{X,Y}(x, y)$ (discr.)
- **Marginal pdf:** $f_X(x) = \int f_{X,Y}(x, y)\, dy$.
- **Independent r.v.'s:** $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)$.
- **Expected sum :** $\mathbb{E}(aX + bY) = a\, \mathbb{E}X + b\, \mathbb{E}Y$.
- **Expected product** (if X, Y independent): $\mathbb{E}[X\, Y] = \mathbb{E}X\, \mathbb{E}Y$.
- **Expectations of a function:** $\mathbb{E}g(X, Y) = \int\int g(x, y)\, f_{X,Y}(x, y)\, dx\, dy$ (cont.) $\mathbb{E}g(X, Y) = \sum\sum g(x, y)\, f_{X,Y}(x, y)$ (cont.)
- **Covariance:** $\mathrm{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$.
- **Properties of Var and Cov**:

$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
$\mathrm{cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$.
$\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
$\mathrm{cov}(aX + bY, Z) = a\, \mathrm{cov}(X, Z) + b\, \mathrm{cov}(Y, Z)$.
$\mathrm{cov}(X, X) = \mathrm{Var}(X)$.
$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\, \mathrm{cov}(X, Y)$.
X and Y independent $\implies \mathrm{cov}(X, Y) = 0$.

- **Conditional pdf:** If $f_X(x) > 0$, $f_{Y \,|\, X}(y \,|\, x) := \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}$.

- The corresponding **conditional expectation**: $\mathbb{E}[Y \mid X = x] = \int y\, f_{Y \mid X}(y \mid x) dy$.
- $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$
- **Moment Generating Function (MGF)**: When it exists, for $t \in I \subset \mathbb{R}$, $M(t) = \mathbb{E}\, \mathrm{e}^{tX} = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\, dx$.
- **MGFs for various distributions:**

$\mathsf{Bin}(n,p)$	$(1-p+pe^s)^n$	$s \in \mathbb{R}$
$\mathsf{Poi}(\lambda)$	$\exp(\lambda(e^s-1))$	$s \in \mathbb{R}$
$\mathcal{U}(a,b)$	$\frac{\mathrm{e}^{bs}-\mathrm{e}^{as}}{t(b-a)}$	$s \in \mathbb{R}$
$\mathsf{Exp}(\lambda)$	$\left(\frac{\lambda}{\lambda-s}\right)$	$s < \lambda$
$\mathcal{N}(\mu,\sigma^2)$	$\mathrm{e}^{s\mu+\sigma^2 s^2/2}$	$s \in \mathbb{R}$

- **Moment property**: $\mathbb{E}X^n = M^{(n)}(0)$.
- $M_{X+Y}(t) = M_X(t)\, M_Y(t)$, $\forall t$, if X, Y independent.
- If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ (independent), then $a + \sum_{i=1}^n b_i X_i \sim \mathcal{N}\left(a + \sum_{i=1}^n b_i\, \mu_i,\ \sum_{i=1}^n b_i^2\, \sigma_i^2\right)$.

Statistics

Tests and Confidence Intervals Based on Standard Errors

- Test statistic: $\frac{\text{estimate}-\text{hypothesised}}{\text{se(estimate)}}$.
- Confidence interval: estimate $\pm$ (critical value) $\times$ se(estimate).
- $se(\bar{x}) = \frac{s}{\sqrt{n}}$; $\quad \mathrm{df} = n-1$
- $se(\bar{x}-\bar{y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$; $\quad \mathrm{df} = \min(n_x - 1, n_y - 1)$
- $se(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- $se(\widehat{p}_x - \widehat{p}_y) = \sqrt{\frac{\widehat{p}_x(1-\widehat{p}_x)}{n_x} + \frac{\widehat{p}_y(1-\widehat{p}_y)}{n_y}}$
- Pooled version for hypothesis testing $se(\widehat{p}_x - \widehat{p}_y) = \sqrt{\widehat{p}(1-\widehat{p})(1/n_x + (1/n_y)}$
- Use t-distribution for means, correlation and regression. Use normal distribution for proportions.

Chi-squared test

- expected count $= \frac{(\text{row total}) \times (\text{column total})}{\text{overall total}}$.
- $X^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}$
- degrees of freedom $= (\#\text{rows} - 1) \times (\#\text{columns} - 1)$.

Linear regression

- $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$, $\quad \mathbf{X}$ is a $n \times p$ matrix
- estimator $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- $\frac{\widehat{\beta}_i - \beta_i}{s.e.(\widehat{\beta}_i)} \sim t_{n-p}$
- $\mathrm{Cov}(\widehat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- $s^2 = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\beta})^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})}{n-p}$

ANOVA

- $DF(Factor) = d-1, \quad DF(Total) = n-1$
- $MS = \frac{SS}{DF}, \quad F = \frac{MSF}{MSE}$

Other Mathematical Formulas

- Factorial. $n! = n\,(n-1)\,(n-2)\cdots 1$. Gives the number of *permutations* (orderings) of $\{1, \ldots, n\}$.
- Binomial coefficient. $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$. Gives the number *combinations* (no order) of k different numbers from $\{1, \ldots, n\}$.
- Newton's binomial theorem: $(a+b)^n = \sum_{k=0}^n a^k\, b^{n-k}$.
- Geometric sum: $1 + a + a^2 + \cdots + a^n = \frac{1-a^{n+1}}{1-a}$ $(a \neq 1)$. If $|a| < 1$ then $1 + a + a^2 + \cdots = \frac{1}{1-a}$.
- Logarithms:
  1. $\log(x\,y) = \log x + \log y$.
  2. $\mathrm{e}^{\log x} = x$.
- Exponential:
  1. $\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.
  2. $\mathrm{e}^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$.
  3. $\mathrm{e}^{x+y} = \mathrm{e}^x\, \mathrm{e}^y$.
- Differentiation:
  1. $(f+g)' = f' + g'$
  2. $(fg)' = f'g + fg'$
  3. $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$
  4. $\frac{d}{dx}x^n = n\,x^{n-1}$
  5. $\frac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$
  6. $\frac{d}{dx}\log(x) = \frac{1}{x}$
- Chain rule: $(f(g(x)))' = f'(g(x))\, g'(x)$.
- Integration: $\int_a^b f(x)\, dx = [F(x)]_a^b = F(b) - F(a)$, where $F' = f$.
- Integration by parts: $\int_a^b f(x)\, G(x)\, dx = [F(x)\, G(x)]_a^b - \int_a^b F(x)\, g(x)\, dx$. (Here $F' = f$ and $G' = g$.)