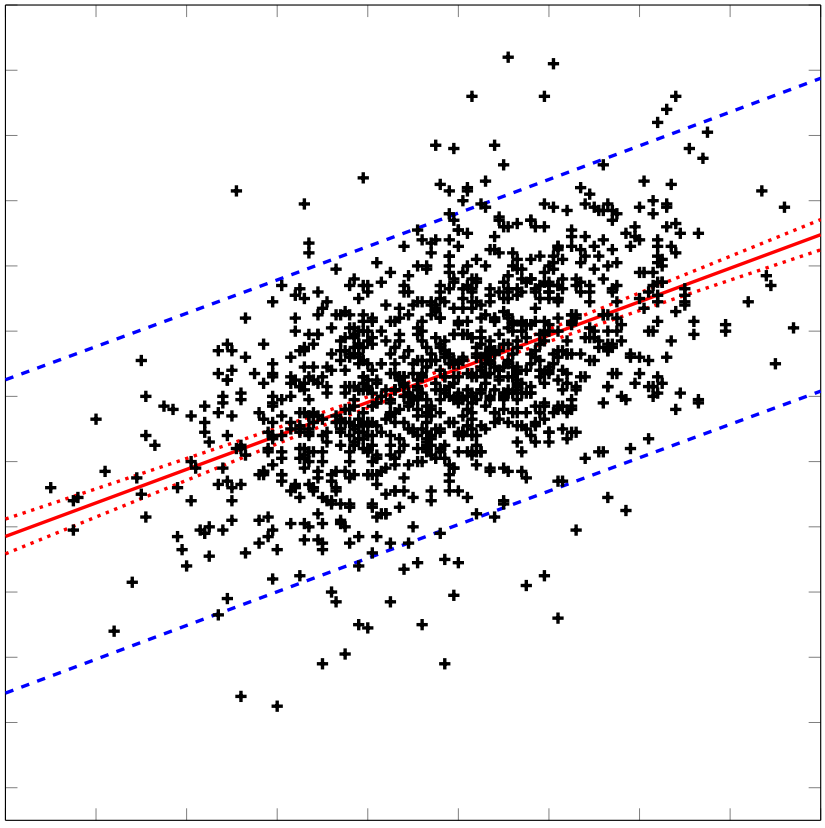
Probability Models and Data Analysis for Engineering (STAT2203)

Applied Probability and Statistics (STAT7203)



Ross McVinish

School of Mathematics and Physics

Preface

These notes are intended as a supplement to, and not a replacement for, lectures. Some material is adapted from [1, 2, 3, 4, 5].

These notes are a revised version of the notes from previous versions of this course by Thomas Taimre, Brendan Patch and Mitchell Griggs.

Ross McVinish *2019*

Acronyms and Abbreviations

ANOVA	Analysis of Variance
cdf	Cumulative distribution function
CI	Confidence Interval
CLT	Central Limit Theorem
Corr	Correlation
Cov	Covariance
iid	Independent and identically distributed
KS	Kolmogorov–Smirnov
LLN	Law of Large Numbers
MLE	Maximum Likelihood Estimate / Estimator / Estimation
mgf	Moment generating function
ODE	Ordinary differential equation
pdf	Probability density function
pgf	Probability generating function
pmf	Probability mass function
rv	random variable
Var	Variance

List of Symbols

$\gg$	much greater than
$\approx$	is approximately
$\propto$	is proportional to
$\sim$	is distributed as
$\overset{\text{iid}}{\sim}$, $\sim_{\text{iid}}$	are independent and identically distributed as
$\overset{\text{approx.}}{\sim}$	is approximately distributed as
∇f	gradient of f
$A^\top$, $\mathbf{x}^\top$	transpose of matrix A or vector $\mathbf{x}$
$\det(A)$	determinant of matrix A
A, B, C	events (sets) or matrices
d, ∂	differential symbol
$\mathbb{E}$	expectation
e	the number $2.71828\ldots$
i	the square root of -1
$\mathrm{I}_A, \mathrm{I}\{A\}$	indicator function of event A
ln	(natural) logarithm
$\mathbb{N}$	set of natural numbers $\{0, 1, \ldots\}$
Ω, Ω_X	sample space
φ	pdf of the standard Normal distribution
Φ	cdf of the standard Normal distribution
$\mathbb{P}$	probability measure
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	the real line = one-dimensional Euclidean space
$\mathbb{R}_+$	positive real line: $[0, \infty)$
$\mathbb{R}^n$	n-dimensional Euclidean space
$\mathbf{x}, \mathbf{y}$	vectors
$\mathbf{X}, \mathbf{Y}$	random vectors
$\mathbb{Z}$	set of integers $\{\ldots, -1, 0, 1, \ldots\}$
z_γ	the γ-quantile of the standard Normal distribution

Standard Normal distribution

	Second decimal place of z									
z	0	1	2	3	4	5	6	7	8	9
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3.0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.3										

This table gives $P(Z \geq z)$ for $Z \sim \text{Normal}(0,1)$. Critical values of the Normal distribution, the z^* values such that $P(Z \geq z^*) = p$ for a particular p, can be found from the ∞ row of the t-distribution Table.

Critical values of Student's T distribution

	Probability p								
df	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6	3183.1
2	0.816	1.886	2.920	4.303	6.965	9.925	22.33	31.60	70.70
3	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92	22.20
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610	13.03
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869	9.678
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781	6.010
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587	5.694
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437	5.453
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318	5.263
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221	5.111
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140	4.985
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015	4.791
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965	4.714
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922	4.648
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883	4.590
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850	4.539
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819	4.493
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792	4.452
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768	4.415
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745	4.382
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725	4.352
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707	4.324
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690	4.299
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674	4.275
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659	4.254
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646	4.234
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551	4.094
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496	4.014
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460	3.962
70	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435	3.926
80	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416	3.899
90	0.677	1.291	1.662	1.987	2.368	2.632	3.183	3.402	3.878
100	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.390	3.862
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

This table gives t^* such that $P(T \geq t^*) = p$, where $T \sim$ Student(df).

χ^2 distribution

	Probability p								
df	0.975	0.95	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	0.001	0.004	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	0.051	0.103	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	0.216	0.352	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	0.484	0.711	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	0.831	1.145	6.626	9.236	11.07	12.83	15.09	16.75	20.52
6	1.237	1.635	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	1.690	2.167	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	2.180	2.733	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	2.700	3.325	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	3.247	3.940	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	3.816	4.575	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	4.404	5.226	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	5.009	5.892	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	5.629	6.571	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	6.262	7.261	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	6.908	7.962	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	7.564	8.672	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	8.231	9.390	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	8.907	10.12	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	9.591	10.85	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	10.28	11.59	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	10.98	12.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	11.69	13.09	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	12.40	13.85	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	13.12	14.61	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	13.84	15.38	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	14.57	16.15	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	15.31	16.93	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	16.05	17.71	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	16.79	18.49	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	24.43	26.51	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	32.36	34.76	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	40.48	43.19	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	48.76	51.74	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	57.15	60.39	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	65.65	69.13	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	74.22	77.93	109.1	118.5	124.3	129.6	135.8	140.2	149.4

This table gives x^* such that $P(X^2 \geq x^*) = p$, where $X^2 \sim \chi^2(\text{df})$.

Summary of Formulas

1. **Sum rule:** $\boxed{\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)}$, when $A_1, A_2, \ldots$ are disjoint.
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
4. **Cdf** of X: $F(x) = \mathbb{P}(X \leqslant x),\ x \in \mathbb{R}$.
5. **Pmf** of X: (discrete r.v.) $f(x) = \mathbb{P}(X = x)$.
6. **Pdf** of X: (continuous r.v.) $f(x) = F'(x)$.
7. For a discrete r.v. X: $\boxed{\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).}$
8. For a continuous r.v. X with pdf f: $\boxed{\mathbb{P}(X \in B) = \int_B f(x)\,dx.}$
9. In particular (continuous), $F(x) = \int_{-\infty}^{x} f(u)\,du$.
10. Similar results 7-8 hold for random vectors, e.g. $\mathbb{P}((X,Y) \in B) = \iint_B f_{X,Y}(x,y)\,dx\,dy$.
11. Marginal from joint pdf: $f_X(x) = \int f_{X,Y}(x,y)\,dy$.
12. **Important discrete distributions**:

Distr.	pmf	$x \in$
Bernoulli(p)	$p^x(1-p)^{1-x}$	$\{0,1\}$
Binomial(n,p)	$\binom{n}{x} p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$
Poisson(λ)	$\mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!}$	$\{0,1,\ldots\}$
Geometric(p)	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$

13. **Important continuous distributions**:

Distr.	pdf	$x \in$
Uniform$[a,b]$	$\frac{1}{b-a}$	$[a,b]$
Exp(λ)	$\lambda\,\mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$
Gamma(α,λ)	$\frac{\lambda^\alpha x^{\alpha-1}\mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$
Normal(μ,σ^2)	$\frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$

14. **Conditional probability:** $\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.
15. **Law of total probability**: $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\,\mathbb{P}(B_i)$,
with $B_1, B_2, \ldots, B_n$ a partition of Ω.
16. **Bayes' Rule:** $\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\,\mathbb{P}(A|B_j)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\,\mathbb{P}(A|B_i)}$.
17. **Product rule**:
$\boxed{\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}).}$
18. **Memoryless property** (Exp and Geometric distribution): $\mathbb{P}(X > s+t \,|\, X > s) = \mathbb{P}(X > t)$, $\forall s,t$.
19. **Independent events**: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
20. **Independent r.v.'s:** (discrete)
$\boxed{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} \mathbb{P}(X_k = x_k)\ .}$
21. **Independent r.v.'s:** (continuous)
$\boxed{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)\ .}$
22. **Expectation** (discr.): $\mathbb{E}\,X = \sum_x x\,\mathbb{P}(X = x)$.
23. (of function) $\mathbb{E}\ g(X) = \sum_x g(x)\,\mathbb{P}(X = x)$.
24. **Expectation** (cont.): $\mathbb{E}\,X = \int x\,f(x)\,dx$.
25. (of function) $\mathbb{E}\ g(X) = \int g(x) f(x)\,dx$,
26. Similar results 22–25 hold for random vectors.
27. **Expected sum :** $\mathbb{E}\,(aX + bY) = a\,\mathbb{E}\,X + b\,\mathbb{E}\,Y$.
28. **Expected product** (only if X, Y independent): $\mathbb{E}\,[X\,Y] = \mathbb{E}\,X\,\mathbb{E}\,Y$.
29. **Markov inequality:** $\mathbb{P}(X \geqslant x) \leqslant \frac{\mathbb{E}\,X}{x}$.
30. $\mathbb{E}\,X$ **and** $\mathbf{Var}(X)$ **for various distributions**:

	$\mathbb{E}\,X$	$\mathrm{Var}(X)$
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	λ	λ
Uniform(a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α,λ)	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Normal(μ,σ^2)	μ	σ^2

31. n**-th moment:** $\mathbb{E}\,X^n$.
32. **Covariance:** $\mathrm{cov}(X,Y) = \mathbb{E}\,(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)$.
33. **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 $\Longrightarrow \mathrm{cov}(X,Y) = 0$.

34. **Moment Generating Function (MGF)**: For discrete distributions, $M(s) = \mathbb{E}\ \mathrm{e}^{sX} = \sum_{n=0}^{\infty} \mathrm{e}^{sn}\mathbb{P}(N = n)$. For continuous distributions, $M(s) = \mathbb{E}\ \mathrm{e}^{sX} = \int_{-\infty}^{\infty} \mathrm{e}^{sx} f(x)\,dx$,
$s \in I \subset \mathbb{R}$, for r.v.'s X for which all moments exist.
35. **MGFs for various distributions:**

Bernoulli(p)	$1 - p + p\,\mathrm{e}^s$
Binomial(n,p)	$(1 - p + p\,\mathrm{e}^s)^n$
Geometric(p)	$\frac{p\,\mathrm{e}^s}{1-(1-p)\,\mathrm{e}^s}$
Poisson(λ)	$\mathrm{e}^{-\lambda(1-z)}$
Uniform(a,b)	$\frac{\mathrm{e}^{bs} - \mathrm{e}^{as}}{s(b-a)}$
Gamma(α,λ)	$\left(\frac{\lambda}{\lambda - s}\right)^\alpha$
Normal(μ,σ^2)	$\mathrm{e}^{s\mu + \sigma^2 s^2/2}$

36. **Laplace transform**: $L(s) := \mathbb{E}\ \mathrm{e}^{-sX}, \quad s \geqslant 0$,
for *positive* r.v.'s.
37. **Moment property**: $\mathbb{E}\,X^n = M^{(n)}(0)$.
38. $M_{X+Y}(t) = M_X(t)\,M_Y(t)$, $\forall t$, if X, Y independent.
39. If $X_i \sim$ Normal(μ_i, σ_i^2), $i = 1,2,\ldots,n$ (independent), then $a + \sum_{i=1}^{n} b_i\,X_i \sim$ Normal $\left(a + \sum_{i=1}^{n} b_i\,\mu_i,\ \sum_{i=1}^{n} b_i^2\,\sigma_i^2\right)$.
40. **Conditional pmf/pdf**

$$f_{Y\,|\,X}(y\,|\,x) := \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}.$$

41. The corresponding **conditional expectation** (discrete case): $\mathbb{E}\,[Y \,|\, X = x] = \sum_y y\,f_{Y\,|\,X}(y\,|\,x)$.

42. **Linear transformation**: $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}}(A^{-1}\boldsymbol{z})}{|A|}$.

43. **General transformation**: $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \dfrac{f_{\boldsymbol{X}}(\mathbf{x})}{|J_{\mathbf{x}}(g)|}$, with $\mathbf{x} = g^{-1}(\boldsymbol{z})$, where $|J_{\mathbf{x}}(g)|$ is the Jacobian of g evaluated at $\boldsymbol{x}$.

44. Pdf of the **multivariate normal** distribution:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}\ .$$

Σ is the covariance matrix, and $\boldsymbol{\mu}$ the mean vector.

45. If $\mathbf{X}$ is a column vector with independent $\mathsf{Normal}(0,1)$ components, and B is a matrix with $\Sigma = BB^T$ (such a B can always be found), then $\mathbf{Z} = \boldsymbol{\mu} + B\mathbf{X}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

46. Pdf of the **multivariate Normal** distribution:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}\ .$$

Σ is the covariance matrix, and $\boldsymbol{\mu}$ the mean vector.

47. If $\mathbf{X}$ has a multivariate Normal distribution $\mathsf{Normal}(\boldsymbol{\mu}, \Sigma)$ (dimension n) and $\mathbf{Y} = \boldsymbol{a} + B\mathbf{X}$ (dimension $m \leqslant n$), then $\mathbf{Y} \sim \mathsf{Normal}(\boldsymbol{a} + B\boldsymbol{\mu}, B\Sigma B^T)$.

48. **Central Limit Theorem**:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x),$$

where Φ is the cdf of the standard Normal distribution.

49. **Normal Approximation to Binomial:** If $X \sim \mathsf{Binomial}(n,p)$, then, for large n, $\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k)$, where $Y \sim \mathsf{Normal}(np, np(1-p))$.

Statistics

Tests and Confidence Intervals Based on Standard Errors

- Test statistic: $\dfrac{\text{estimate}-\text{hypothesised}}{\text{se(estimate)}}$.
- Confidence interval:
  estimate $\pm$ (critical value) $\times$ se(estimate).
- $se(\bar{x}) = \frac{s}{\sqrt{n}}$
- $se(\bar{x} - \bar{y}) = s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$
- (pooled sample variance) $s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}$.
- $se(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- $se(\widehat{p}_x - \widehat{p}_y) = \sqrt{\frac{\widehat{p}_x(1-\widehat{p}_x)}{n_x} + \frac{\widehat{p}_y(1-\widehat{p}_y)}{n_y}}$
- Use t-distribution for means, correlation and regression. Use normal distribution for proportions.

Chi-squared test

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

Linear regression

- $\mathbf{Y} \sim \mathsf{Normal}(\mathbf{X}\beta, \sigma^2 I)$
- estimator $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- $\mathrm{Cov}(\widehat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- $s^2 = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\beta})^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})}{n-p}$

Other Mathematical Formulas

1. Factorial. $n! = n\,(n-1)\,(n-2)\cdots 1$. Gives the number of *permutations* (orderings) of $\{1,\ldots,n\}$.
2. Binomial coefficient. $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$. Gives the number *combinations* (no order) of k different numbers from $\{1,\ldots,n\}$.
3. Newton's binomial theorem: $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k\,b^{n-k}$.
4. Geometric sum: $1 + a + a^2 + \cdots + a^n = \frac{1-a^{n+1}}{1-a}$ $(a \neq 1)$. If $|a| < 1$ then $1 + a + a^2 + \cdots = \frac{1}{1-a}$.
5. Logarithms:
   (a) $\ln(x\,y) = \ln x + \ln y$.
   (b) $\mathrm{e}^{\ln x} = x$.
6. Exponential:
   (a) $\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.
   (b) $\mathrm{e}^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$.
   (c) $\mathrm{e}^{x+y} = \mathrm{e}^x\,\mathrm{e}^y$.
7. Differentiation:
   (a) $(f+g)' = f' + g'$,
   (b) $(fg)' = f'g + fg'$,
   (c) $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$.
   (d) $\frac{d}{dx}x^n = n\,x^{n-1}$.
   (e) $\frac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$.
   (f) $\frac{d}{dx}\ln(x) = \frac{1}{x}$.
8. Chain rule: $(f(g(x)))' = f'(g(x))\,g'(x)$.
9. Integration: $\int_a^b f(x)\,dx = [F(x)]_a^b = F(b) - F(a)$, where $F' = f$.
10. 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 $G' = f$.)
11. Jacobian: Let $\mathbf{x} = (x_1,\ldots,x_n)$ be an n-dimensional vector, and $g(\mathbf{x}) = (g_1(\mathbf{x}),\ldots,g_n(\mathbf{x}))$ be a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. The *matrix of Jacobi* is the matrix of partial derivatives: $(\partial g_i/\partial x_j)$. The corresponding determinant is called the *Jacobian*. In the neighbourhood of any fixed point, g behaves like a *linear transformation* specified by the matrix of Jacobi at that point.
12. Γ function: $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\,\mathrm{e}^{-u}\,du$, $\alpha > 0$. $\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$, for $\alpha \in \mathbb{R}_+$. $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$. $\Gamma(1/2) = \sqrt{\pi}$.

Contents

1

Introduction

During the years that I've been preparing Volume 4B, I've often run across basic techniques of probability theory that I would have put into Section 1.2 of Volume 1 had I been clairvoyant enough to anticipate them in the 1960s ...

— Donald Knuth preface to Volume 4B of The Art of Computer Progamming.

By the end of this chapter you should:

- Know the aims of this course.

Please consult the Electronic Course Profile for details concerning assessment.

A **random experiment** is a process whose outcome cannot be determined in advance.

Examples of random experiments:

- The number of collisions in a hash table.
- Time for a bug in a program to be found/reported.
- The number of comparisons required by the quicksort algorithm to sort a list of items.

To handle *randomness*, we need **models** for random experiments.

Dogfight

Suppose that Alice, Bob, and Carol are fighting in an air battle.

- In each round, each survivor fires one shot. Alice fires first, then Carol fires, and then Bob fires.
- Anyone hit drops out of the battle immediately.

- On any shot aimed at an opponent:
  - Alice hits with probability $\frac{2}{5}$. (Meaning on average 2 out of every 5 shots Alice fires will hit their target.)
  - Bob hits with probability $\frac{1}{2}$.
  - Carol never misses.

Question: Where should Alice fire in the first round?

Answer:

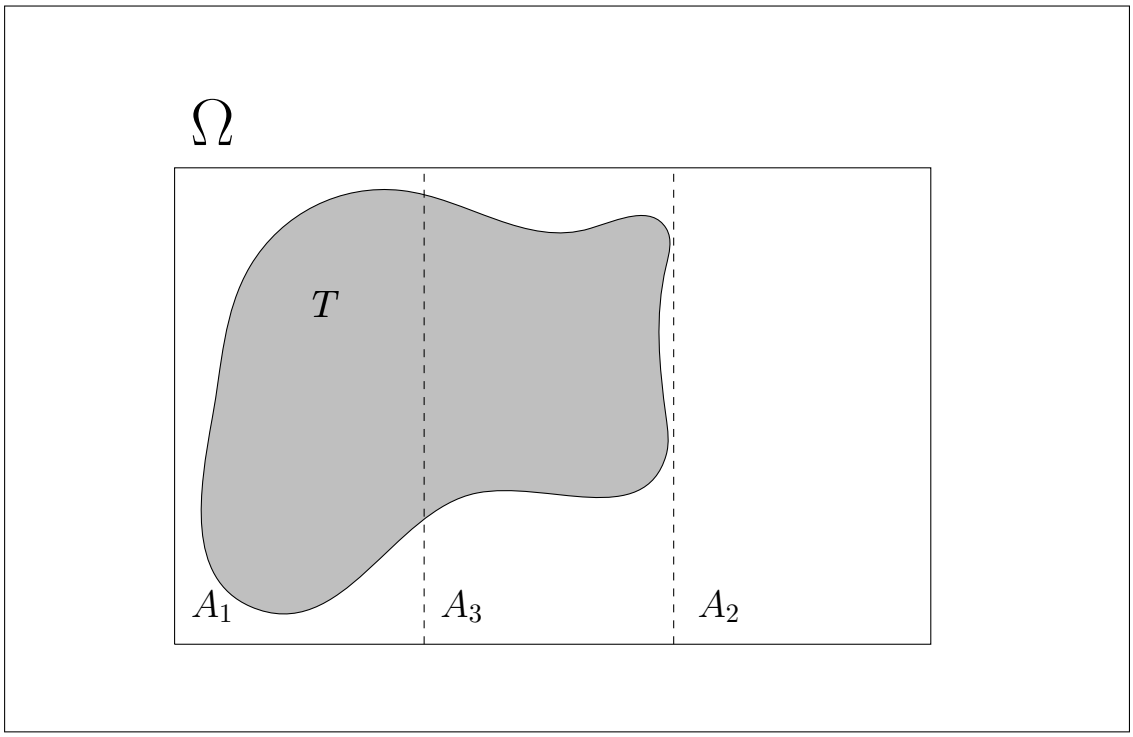
Note that if Alice fires at Bob and hits him, then Alice will immediately be shot by Carol and lose. So in the first round Carol is the only target to consider.

First, suppose that Alice was to fire at Carol and hits her (which occurs with probability $\frac{2}{5}$). Alice would then be Bob's target, and the battle would continue with shots alternating between Alice and Bob until a hit is scored. Define the **event**

- $T \equiv$ Bob wins the one-on-one battle with Alice.

We also need to consider the **events**:

- $A_1 \equiv$
- $A_2 \equiv$
- $A_3 \equiv$



If A_3 occurs, then the next round begins under the same conditions; hence,

$$\mathbb{P}(T \mid A_3) =$$

here the function $\mathbb{P}$ takes an event (e.g. $T \mid A_3$) and gives its probability. When we write $T \mid A_3$ we are supposing that the event A_3 has already happened, and we are now interested in the event T *conditional* on the fact that A_3 has occurred. Also,

$$\mathbb{P}(T \mid A_1) = \qquad \text{and} \quad \mathbb{P}(T \mid A_2) =$$

because if Bob hits with his first shot then Alice is out of the game and Bob wins, but if Bob misses with his first shot and then Alice hits with her first shot then Bob is out of the game.

So

$$\mathbb{P}(T) = \mathbb{P}(T \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(T \mid A_2)\mathbb{P}(A_2) + \mathbb{P}(T \mid A_3)\mathbb{P}(A_3)$$

$$=$$

which implies

$$=$$

$$=$$

Second, suppose that Alice was to fire at Carol and misses her. Then Carol will certainly fire at Bob because $\frac{2}{5} < \frac{1}{2}$ and Carol is a rational person. Since Carol is such a good shot, she will 'almost surely' hit Bob. So in this case Alice only wins if Alice's second shot hits Carol in the second round, which happens with probability $\frac{2}{5}$.

Hence, missing Carol (on purpose) gives Alice a better chance since

$$\underbrace{1 - \mathbb{P}(T)}_{\text{Alice wins against Bob}} = 1 - \frac{5}{7} = \frac{2}{7} < \frac{2}{5}\,.$$

In summary, since Bob is a sufficiently better shot than Alice, Alice should let Carol take out Bob and hope for a good outcome on her single shot against Carol.

This example is adapted from [5, p. 68].

Figure 1.1: From `www.xkcd.com/1323/`

Software effort estimation

Cost-estimation is a difficult problem in software engineering. A basic model that relates the person-hours a project requires and a measure of the complexity of the project is

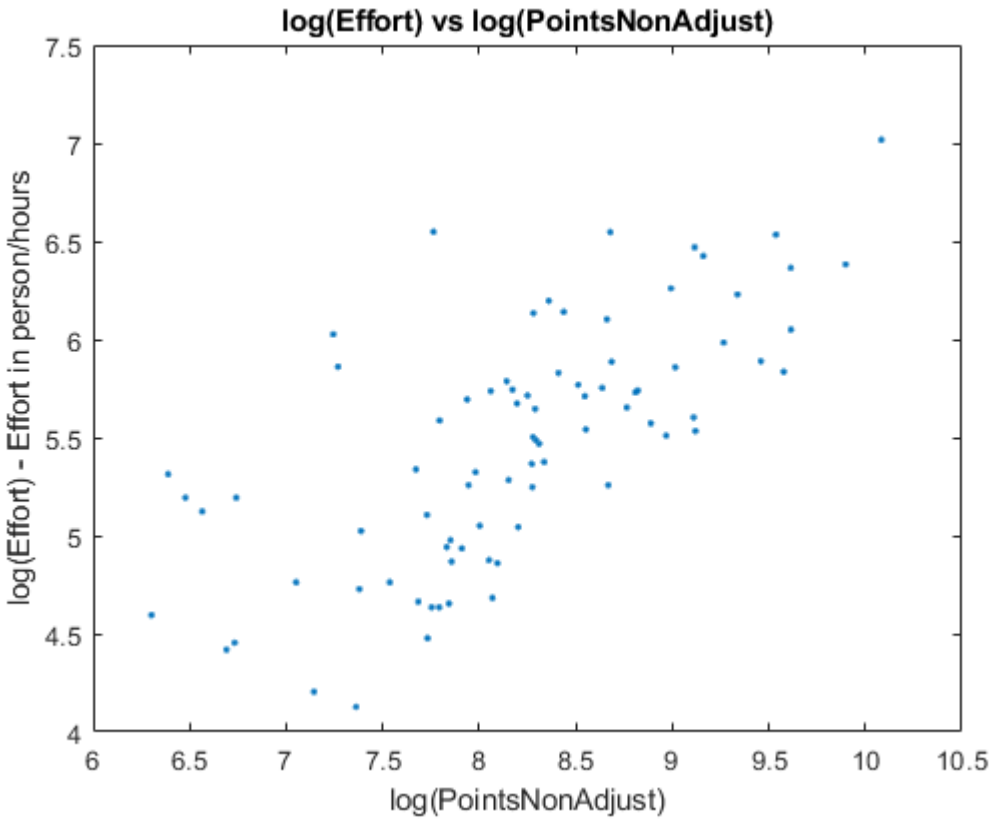
$$\text{Effort} = a \times (\text{Project complexity})^b,$$

where a and b are parameters that can be estimated from historical data. For the purpose of this demonstration we take the non-adjusted function point technique as the measure of complexity. Jean-Marc Desharnais[1] surveyed 10 organisations on 81 management information systems development projects completed between 1983 and 1988. The data is presented in the scatterplot below on the log-log scale.

Question: How can we estimate a and b from this data?

```
desh = readtable('desharnais.xlsx');
desh.logEffort = log(desh.Effort);
desh.logPNA = log(desh.PointsNonAdjust);
plot(desh.logEffort,desh.logPNA,'.')
title('log(Effort) vs log(PointsNonAdjust)')
xlabel('log(PointsNonAdjust)')
ylabel('log(Effort) — Effort in person/hours')
```

[1] The file `deshardnais.xlsx` is on Blackboard. The original data from Desharnais's Masters thesis is available from `http://promise.site.uottawa.ca/SERepository/datasets-page.html`.



Notice that log(*Effort*) is roughtly a linear function of the log(*Project Complexity*).

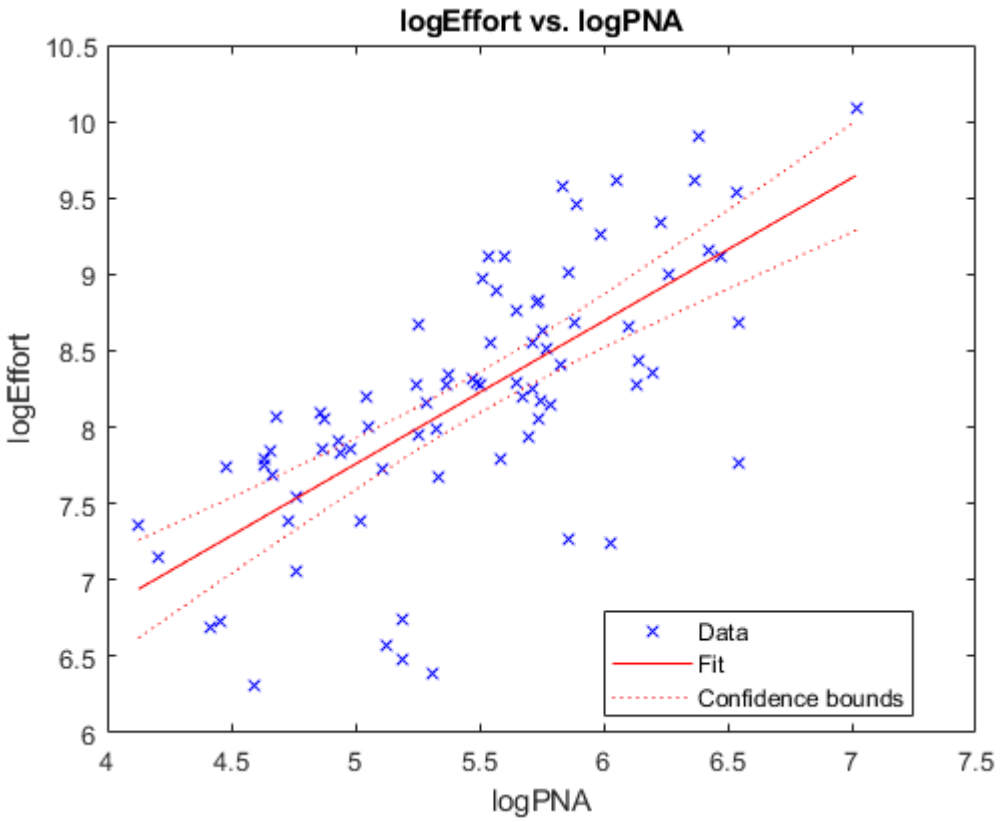
Question: Suppose that a we have a project whose log(*Project Complexity*) is 6. How many person hours do we expect the project to take? How certain would we be of this estimate.

The following code can be used to perform an *ordinary least squares* analysis of this dataset in MATLAB. First place the file 'desharnais.xlsx' in your current working directory, and then execute the following code.

```
1 lm = fitlm(desh, 'logEffort~logPNA')
2 plot(lm)
```

You'll see the following output and figure.

```
Linear regression model:
    logEffort ~ 1 + logPNA

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)      3.0775     0.59842    5.1427    1.9218e-06
    logPNA          0.93612     0.10863    8.6178     5.428e-13


Number of observations: 81, Error degrees of freedom: 79
Root Mean Squared Error: 0.599
R-squared: 0.485,  Adjusted R-Squared 0.478
F-statistic vs. constant model: 74.3, p-value = 5.43e-13
```



Having obtained estimates for a and b, our task is not finished. We need to ask ourselves:

Question: Is the data consistent with the original model?

If the data is not consistent with the model, then our estimates and inferences will be worthless.

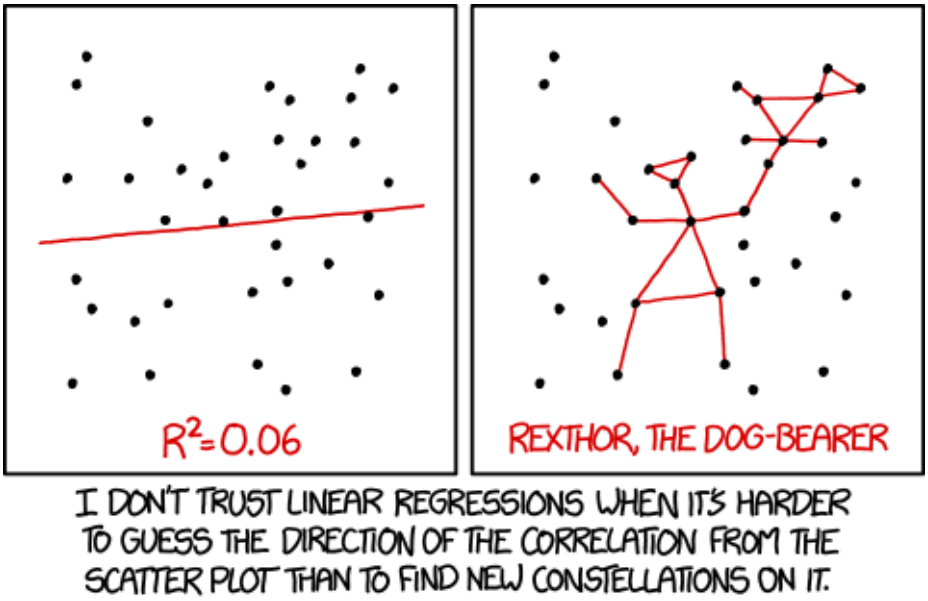


Figure 1.2: From `www.xkcd.com/1725/`

2

Sample Spaces, Events, and Probabilities

By the end of this chapter you should:

- Know what sample spaces and events are.
- Be able to construct events given a sample space.
- Understand the importance of the probability axioms.
- Work with simple discrete probabilities.

Core idea in probability and statistics: **random experiment**.

> *A random experiment is a process whose outcome is not known in advance, but is nevertheless still subject to analysis.*

Examples:

To model any of these examples, we need:

(i) Sample space (written Ω, all possible outcomes).

(ii) Events ($A \subset \Omega$, groups of outcomes).

(iii) Probabilities of events (written $\mathbb{P}(A)$)

Sample spaces and events

Sample space for the experiment "Winner of the 2018 World Cup":

$\Omega =$

Some possible events:

$A =$ winner comes from Asian Football Confederation
$=$

$B =$ Australia wins
$=$

Sample space for the experiment "views of your next Instagram photo":

$\Omega =$

Some possible events:

$A = 10$ views $=$
$B =$ more than 10 views $=$
$C =$ more than 2, but less than 60 views $=$

Definition. The **sample space** Ω of a random experiment is the *set of all possible outcomes of the experiment.*

Definition. An **event** is a *subset of the sample space.* That is, a collection of some possible outcomes of the experiment.

Events will be denoted by capital letters $A, B, C, \ldots$.

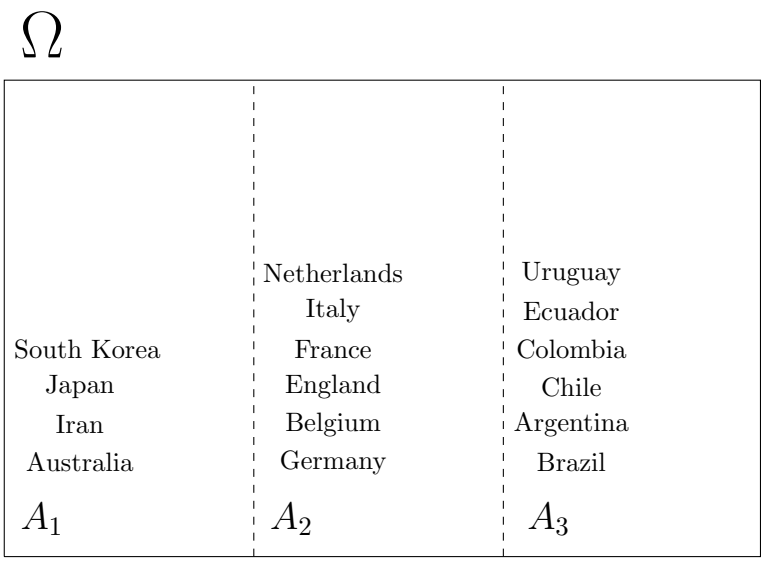
We say event A **occurs** if:

.

When writing down sample spaces make sure you understand the difference between *sets* $\{\ldots\}$ and *vectors* $(\ldots)$:

- *Round* brackets ( ) indicate *order*,
  e.g., $(1, 2, 3) \neq (3, 2, 1)$.
- *Curly* brackets { } indicate *no order*,
  e.g., $\{1, 2, 3\} = \{3, 2, 1\}$.

A sample space with some events marked:



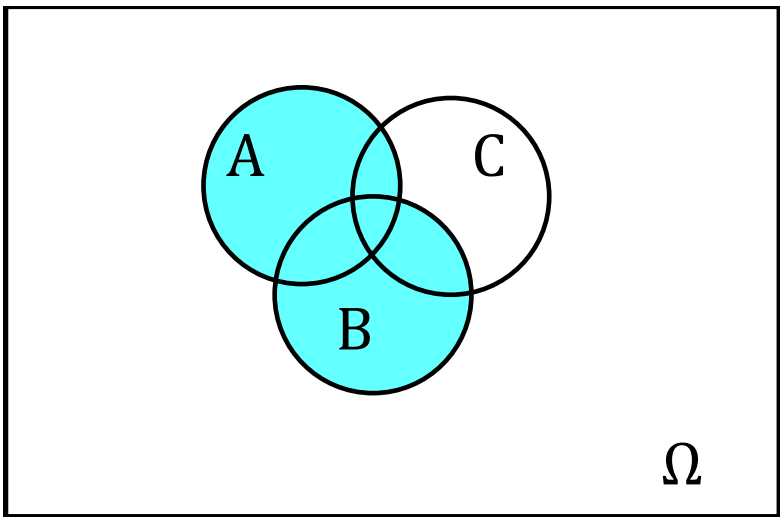
$A_1 = \{\text{South Korea, Iran, Japan, Australia}\} =$
$A_2 = \{\text{Netherlands, Italy, France, England, Belgium, Germany}, \dots\} =$
$A_3 = \{\text{Argentina, Brazil, Chile, Colombia, Ecuador, Uruguay}\} =$
$A_4 = \{\text{Germany, Argentina, Netherlands}\} =$

- The event that there are fewer than 50 defective components in a batch of 1000.
  $A =$
- The event that a machines lives longer than 1000 days,
  $A =$
- The event that between 10 and 20 inclusive hits occur to a web-server, during a specified time interval,
  $A =$
- The event that the sum of two dice is 10 or more (supposing the dice are thrown consecutively):
  $A =$

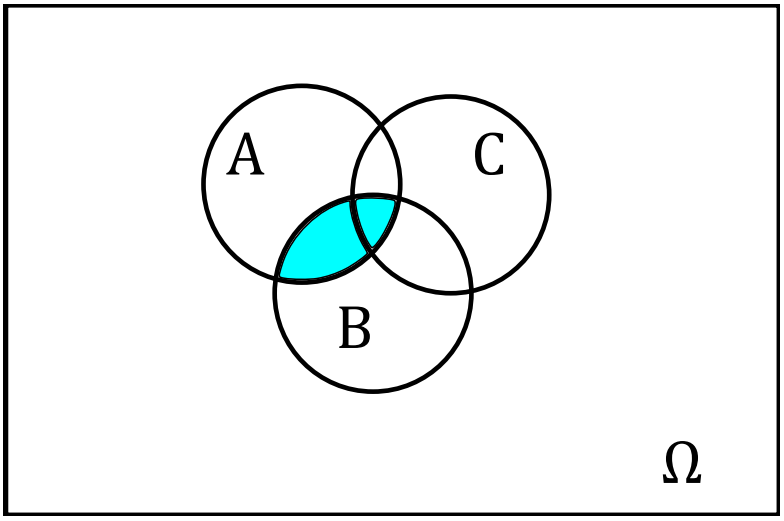
Set operations and events

As events are *sets* of outcomes, usual set operations and definitions apply.

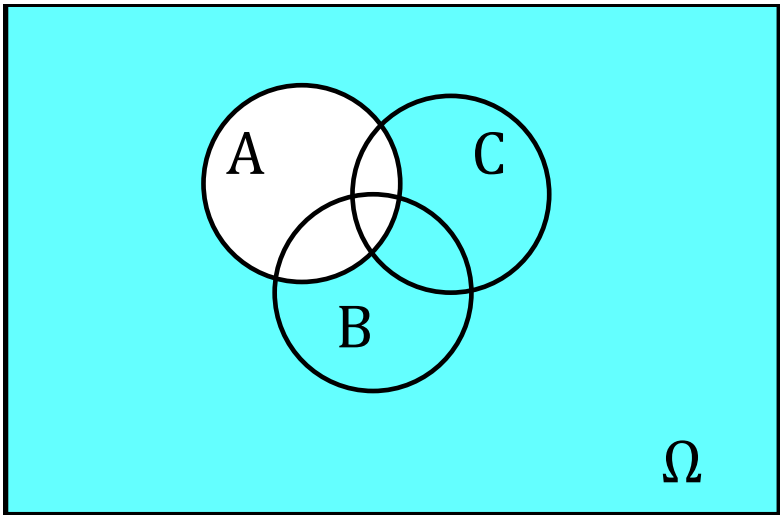
- The event $A \cup B$ (A **union** B) is the event that A *or* B *or both* occur.



- The event $A \cap B$ (A **intersection** B) is the event that A *and* B both occur.



- The event A^c (A **complement** with respect to Ω) is the event that A does not occur.



- Two events A and B are **equal**, $A = B$, if

- Event A **implies** event B, $A \subseteq B$, if A is a subset of B.
- The event containing no outcomes is denoted $\emptyset$ (the **empty** set or impossible event).

Example: Consider the experiment of tossing a standard, six-sided die, whose outcome is an element of the sample space Ω. Let A be the event that the outcome is an even number, and let B be the event that the outcome is an odd number. What are the following sets?

$$\Omega =$$
$$A =$$
$$B =$$

Let $C = \{5, 6\}$ and $D = A \cap C$. What are these next sets?

$$A \cup B =$$
$$A \cap B =$$
$$A^c =$$
$$A \cap C =$$
$$B \cap C =$$
$$C^c =$$

Are the following claims true?

$$D \subseteq C$$
$$C \subseteq A$$
$$D \subseteq A$$
$$D \subseteq B$$

Disjoint Events

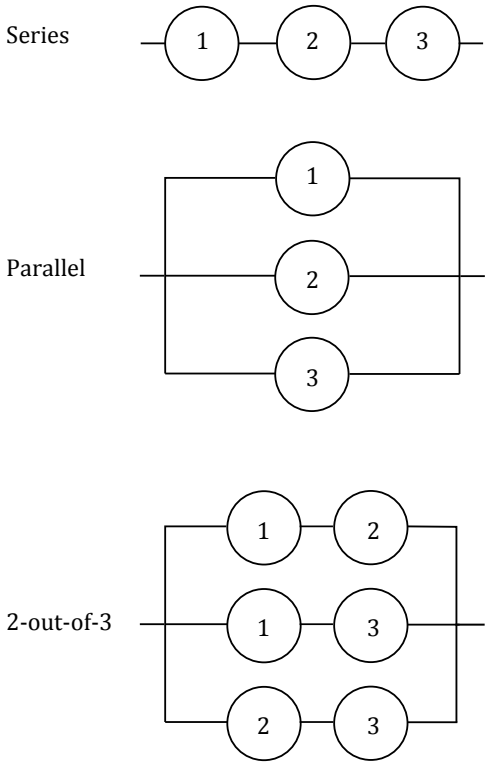
Events $A_1, A_2, \ldots, A_n$ are called **disjoint** if

A sequence $A_1, A_2, \ldots, A_n$ of disjoint events such that their union is the entire sample space Ω is called a **partition** of Ω.

Example: In the above example, $A_1 = \{1, 2\}$, $A_2 = \{3, 4\}$, $A_3 = \{5, 6\}$ are a partition of Ω.

Constructing Events

Example: System Reliability.



Let A_i be the event that *the i-th component is functioning*, $i = 1, 2, 3$. Note that A_i^c is the event that the *i-th* component failed.

D_a = the series system is functioning
=

D_b = the parallel system is functioning
=

D_c = the 2-out-of-3 system is functioning
=

Axioms and Implications

Definition. A **probability** $\mathbb{P}$ is a rule (or function) which assigns a number to each event, and which satisfies the following **axioms** (or properties):

- Axiom 1: $\mathbb{P}(A) \geqslant$.
- Axiom 2: $\mathbb{P}(\Omega) =$.
- Axiom 3: **Sum Rule**: For any disjoint $A_1, A_2, \ldots$

Some **consequences** of the axioms:

- Consequence 1: If $A \subseteq B$ then
- Consequence 2: $\mathbb{P}(\emptyset) =$
- Consequence 3: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- Consequence 4: $0 \leqslant \mathbb{P}(A) \leqslant 1$.
- Consequence 5: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proving these involves cleverly using the axioms, above.

Example: Prove Consequence 3.

$\Omega =$

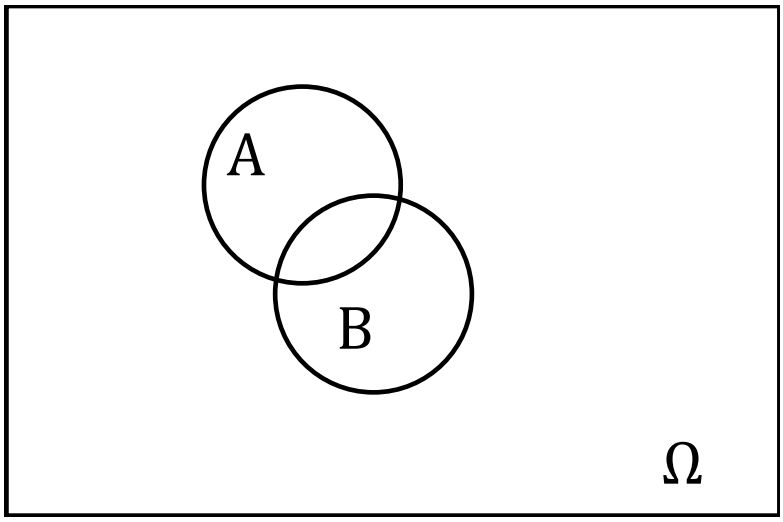
$1 =$ (Axiom 2)

$=$

$=$ (Axiom 3).

Now rearrange this equation.

Question: Visually, what does Consequence 5 mean?



Note that these simple rules of probability are highly similar to those one would use to measure length, area, volume, weight, etc.

Example: Consider a fair, two-sided coin, with sample space $\Omega = \{H, T\}$ (here, the word *fair* means $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$). Write out all possible events and verify that the third probability axiom is satisfied in the situation $A_1 = \{H\}$, $A_2 = \{T\}$.

The events are

$$\mathbb{P}(A_1 \cup A_2) =$$

and

$$\mathbb{P}(A_1) + \mathbb{P}(A_2) =$$

Question: In the above example, $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(A_1) = 1/2$. Is it true that

$$\mathbb{P}(\Omega \cup A_1) = \mathbb{P}(\Omega) + \mathbb{P}(A_1) = 1 + \frac{1}{2}\,?$$

Why/why not?

This is because .

In fact,

Discrete Sample Spaces

If Ω is *countable* it is called a **discrete** sample space; otherwise it is called a **continuous** sample space.

Let Ω be a discrete sample space, e.g. $\Omega = \{a_1, a_2, \ldots, a_n\}$.

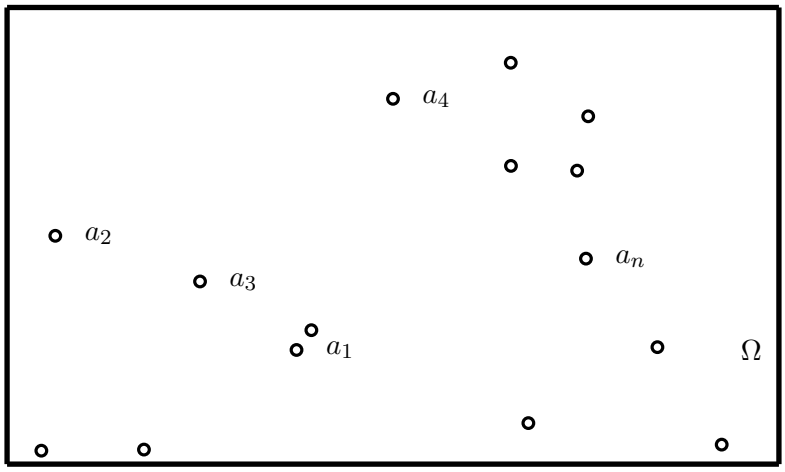


Figure 2.1: A finite sample space $\Omega = \{a_1, a_2, \ldots, a_n\}$.

Let $\mathbb{P}(\{a_i\}) = p_i$, for $i = 1, \ldots, n$, and define

$$\mathbb{P}(A) = \sum_{i:a_i \in A} p_i \ , \quad \text{for all} \ \ A \subset \Omega.$$

Then $\mathbb{P}$ is a probability measure.

Thus we can *specify* $\mathbb{P}$ by specifying only the probabilities of the **elementary events** $\{a_i\}$.

Example. Experiment: throw a fair die.

Sample space:

$$\Omega =$$

Define $\mathbb{P}$ by:

$$\mathbb{P}(A) =$$

This completely specifies/models the experiment. For example, the probability of getting an even number is

$$\mathbb{P}(\{2, 4, 6\}) =$$

Example. We draw two cards from a full deck of 52 cards. What is the probability of drawing at least one Ace?

Give all the cards a number from 1 to 52. Draw the cards one-by-one. Write a possible outcome as (x, y), $x \neq y$.

Each elementary event $\{(x, y)\}$ has the same probability

Let A be the event: "at least one Ace". Then,

$$\mathbb{P}(A) =$$

We need to *count* how many elements are in A. Easier: $|A^c| = 48 \times 47$. Hence,

$$\mathbb{P}(A) =$$

Remark. In many cases, as above, we can choose Ω such that each elementary event is equally likely, i.e. $\mathbb{P}(\{a_i\}) = 1/n$.

This is sometimes known as the *Equally-Likely Principle*, or the *Equilikely Principle*.

Question: What if we choose the two cards *at the same time* (no order). Does that change the model? Does it change the probability?

Example: Consider tossing a fair coin twice (so that you sample from $\{H, T\}$ *with replacement*). What is the probability of getting both one heads and one tails?

- **Model I:** Order recorded:

$$\begin{aligned} \Omega &= \\ A &= \\ \mathbb{P}(A) &= \frac{|A|}{|\Omega|} = \frac{|A|}{2^2} = \frac{2}{4} = \frac{1}{2} \end{aligned}$$

- **Model II:** Order ignored:

$$\begin{aligned} \widetilde{\Omega} &= \{\{H, H\}, \{H, T\}, \{T, T\}\} \\ \widetilde{A} &= \{\{H, T\}\} \\ \widetilde{\mathbb{P}}(\widetilde{A}) &= \frac{1}{2} \end{aligned}$$

Question: Did we apply the Equally-Likely Principle in Model I of the above example? Did we apply it in Model II?

Example: Suppose that we have three balls (labelled 1, 2, and 3) in an urn, and select (*without replacement*) two of the balls. Consider the event of 1 being chosen as one of the two balls.

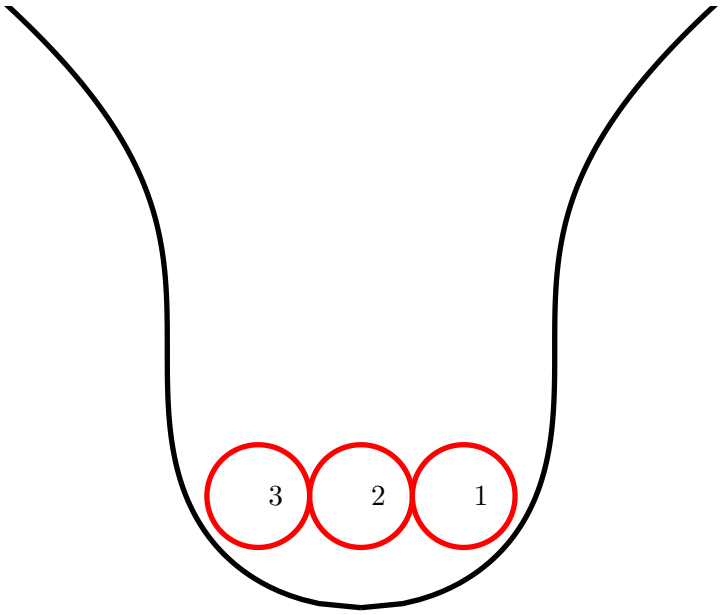


Figure 2.2: Balls 1, 2, and 3 in an urn.

- **Model I:** Order recorded:

$$\begin{aligned}\Omega &= \\ A &= \\ \mathbb{P}(A) &= \frac{|A|}{|\Omega|} = \frac{|A|}{6} = \frac{4}{6} = \frac{2}{3}\end{aligned}$$

- **Model II:** Order ignored:

$$\begin{aligned}\widetilde{\Omega} &= \{\{1,2\},\, \{1,3\}, \{2,3\}\} \\ \widetilde{A} &= \{\{1,2\},\, \{1,3\}\} \\ \widetilde{\mathbb{P}}(\widetilde{A}) &= \frac{|\widetilde{A}|}{|\widetilde{\Omega}|} = \frac{|\widetilde{A}|}{3} = \frac{2}{3}.\end{aligned}$$

Counting

In the previous examples we drew two cards from a full deck, and balls from an urn, *without replacement*. In general, if choosing r objects from a collection of n objects, without replacement, then there are

$$n \times (n-1) \times \cdots \times (n-r+1)$$

ways of doing this. We write ${}^nP_r$ for this quantity. (Check your calculator!) Another way to write this (n permutation r) is

$${}^nP_r = \frac{n!}{(n-r)!}\,,$$

where $n!$ ("n factorial") is given by

$$n! = n \times (n-1) \times \cdots \times 2 \times 1\,.$$

Notice that the *order* of the letters matters.

Example. Two letters are chosen *with replacement* from the word $PING$. What is the sample space? How many outcomes are there?

$$\Omega = \{PP, PI, PN, PG, IP, II, IN, IG, NP, NI, NN, NG, GP, GI,$$
$$|\Omega| =$$

Notice that the *order* of the letters matters.

Example. How many three-letter words can be obtained from the letters in the word $SURFING$? This is sampling without replacement. Does the order matter?

$$|\Omega| =$$

Example. Three people from the set {Joffrey, Balon, Robb, Stannis, Renly} must be chosen to go into a fighting pit. How many different combinations of combatants are there?

$$|\Omega| =$$

Some outcomes are equivalent though, so this is too many! Our event is a union of several outcomes.
Considering the opponents {Robb, Balon, Stannis}. This combination has been counted ${}^3P_3 = 6$ times:

$$RBS, \quad RSB, \quad BRS,$$

If N is the number of ways of choosing opponents, then $N \times {}^3P_3 = {}^5P_3$, so

$$N = {}^5P_3/{}^3P_3 = \frac{5!}{(5-3)!3!} =$$

In general, if choosing r objects from a collection of n objects, without replacement, then the number of combinations is:

$${}^nC_r = \frac{n!}{\qquad\qquad}\qquad .$$

Summary.

	Order matters	Order does not matter
With replacement	n^r	
Without replacement	${}^nP_r = \frac{n!}{(n-r)!}$	${}^nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}$

Matlab code.

```
n = 6;
r = 2;
factorial(n)
ans =
   720
nchoosek(n,r)
ans =
    15
nPr = nchoosek(n,r)*factorial(r)
nPr =
    30
```

Example. How many ways are there to order the letters in the word $INDOOROOPILLY$?

Notice that this situation does not fall into any of the above categories. In general, the number of permutations of n objects with n_1 of type 1, n_2 of type 2, et cetera, is given by

$$\frac{n!}{n_1!n_2!\dots n_k!},$$

where k is the number of types.

Hence, the number of ways to order the letters in $INDOOROOPILLY$ is

Example. Suppose that you have two red balls and three blue balls. How many *distinct* orderings of all balls are there?

3

Conditional Probability and Independent Events

By the end of this chapter you should:

- Understand the rule for conditional probability $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.
- Know what it means for events to be independent.
- Be able to apply Bayes' rule to simple scenarios.

Conditional Probability

How do probabilities *change* when we know *some event $B \subseteq \Omega$ has occurred?*

Suppose B *has occurred*. Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the *relative chance of A occurring* is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$.

This leads to the definition of the **conditional probability** of A given B:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Remark: This only makes sense if $\mathbb{P}(B) > 0$.

Example. We throw two dice (assume consecutively). Given that the *sum of the faces is 10*, what is the probability that *one 6 is cast*?

Let B be the event that *the sum is 10*,

$B =$

Let A be the event that *one 6 is cast*,

$A =$

Then, $A \cap B =$. Since all elementary events are equally likely, we have

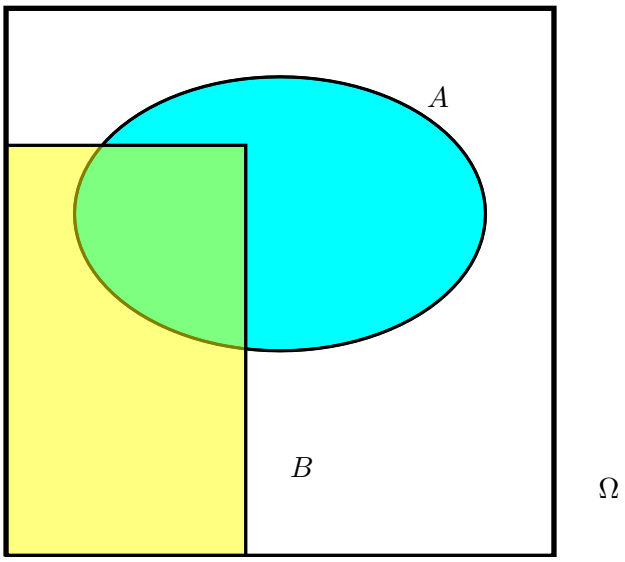


Figure 3.1: The probability of A is . However, if B has occurred then the probability of A *given* B is .

$\mathbb{P}(A \mid B) =$

Compare this with

$\mathbb{P}(A) =$

Question. Morgan (a boy) has exactly one sibling. What is the probability that Morgan has a sister?

Now consider the following arguments:

- Assuming that the sex of siblings is independent, the other sibling is equally likely to be a girl or a boy. The answer is 1/2.
- There are four cases: BB, BG, GB, GG. The last is impossible, and in two of the remaining three cases the sibling is a girl. The answer is 2/3.

Actually this question is poorly posed, and highlights the importance of clearly specifying the model and events of interest since the answer depends on whether we know only that Morgan is one of two children, or if we know that Morgan is (for instance) the older sibling.

Suppose that a couple had two children, where each child was independently, equally likely to be a boy or a girl. Given that one of them is a boy, what is the probability that the other is a girl?

The sample space is

$$\Omega =$$

and the event of interest is $A =$
and we know that the event $B =$
occurs. Hence,

$$\mathbb{P}(A \mid B) =$$

What is the probability that the second (to be born) child is a girl if the first one is a boy?

The sample space is

$$\Omega =$$

the event of interest is $A =$
and we know that the event $B =$
occurs. Hence,

$$\mathbb{P}(A \mid B) =$$

Question: Suppose that A and B are events where both $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Can we write $\mathbb{P}(B \mid A)$ in terms of $\mathbb{P}(A \mid B)$?

$\mathbb{P}(B \mid A) =$ and

$\mathbb{P}(A \mid B) =$ so

, or

$\mathbb{P}(B \mid A) =$.

This is a special case of *Bayes' Rule*, which we will see later in this chapter.

Example. A new COVID-19 rapid test claims that a 99% sensitivity (probability of correctly diagnosing a person with COVID-19) and 98% specificity (probability of correctly diagnosing a person who does not have COVID-19). The probability that a randomly selected person in a particular suburb has COVID-19 is 4%.

Given a randomly selected person has tested positive to COVID-19, what is the probability that they have COVID-19?

Let A be the event that a person tests positive to COVID-19 and let B be the person actually has COVID-19. By the above rule,

$$\mathbb{P}(B \mid A) =$$

Product Rule

By the definition of conditional probability we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A).$$

We can generalize this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$ (abbreviated as $A_1 A_2 \cdots A_n$).

This gives the **product rule** (also called *chain rule*) of probability:

$$\mathbb{P}(A_1 A_2 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_1 A_2 \cdots A_{n-1}).$$

Example. Draw five cards from a full deck of 52 cards. What is the probability of no Ace?

Let A_i be event that the i-th draw is not an ace, $i = 1, 2, 3, 4, 5$.

We are interested in $A := A_1 \cap \cdots \cap A_5$.

We know

$$\mathbb{P}(A_1) = \quad ,$$

$$\mathbb{P}(A_2 \mid A_1) = \quad ,$$

$$\mathbb{P}(A_3 \mid A_1 \cap A_2) = \quad ,$$

$$\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) = \quad , \text{ and}$$

$$\mathbb{P}(A_5 \mid A_1 \cap A_2 \cap A_3 \cap A_4) = \quad .$$

Using the product rule, we see that

$$\mathbb{P}(A) =$$

Exercise. (Birthday problem) What is the probability P that no one in a randomly selected group of $n < 365$ persons shares a birthday with someone else? (Assume 365 equally-likely birthdays.)

$$P =$$

Some approximations of this probability:

n	23	40	57
P	0.4927	0.1088	0.0099

Law of Total Probability

Suppose $B_1, B_2, \ldots, B_n$ is a *partition* of Ω. Then, by the *sum rule*,

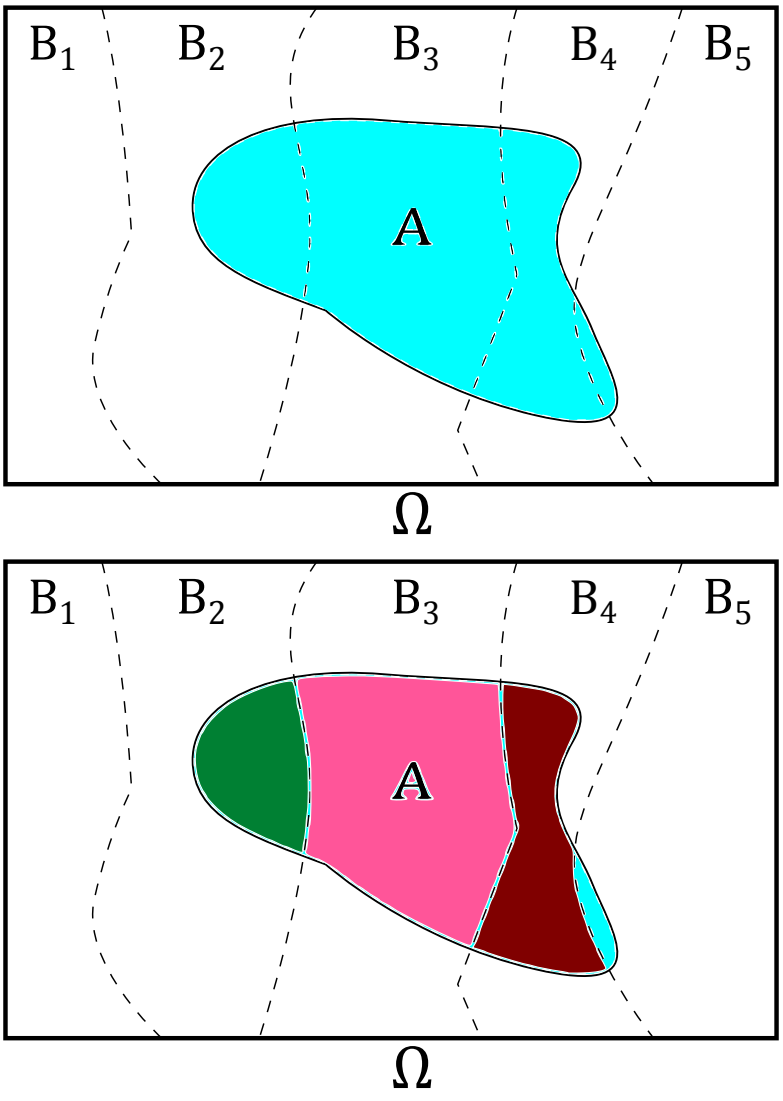
$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) \; .$$

Hence,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)$$

(as long as $\mathbb{P}(B_1) > 0$, $\mathbb{P}(B_2) > 0$, $\ldots$, $\mathbb{P}(B_n) > 0$).

This is called the **Law of Total Probability**.



Example. Draw a card from a full deck of 52 cards. What is the probability it is an Ace?

Let A be the event that the card is an ace. Also, let B_1 be the event that the card is red and B_2 be the event that the card is black. Note that $\Omega = B_1 \cup B_2$ and $B_1 \cap B_2 = \emptyset$ so that $\{B_1, B_2\}$ is a valid partition. Always check this!

$\mathbb{P}(A) =$

Bayes' Rule

Combining the definition of conditional probability with the Law of Total Probability gives the famous rule

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^{n}\mathbb{P}(A|B_i)\mathbb{P}(B_i)} = \frac{\mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}{\mathbb{P}(A)}\,.$$

Example. Email arriving in your inbox is spam 30% of the time. Each of your emails is flagged *correctly* as spam with probability 0.99. However, each of your emails is flagged *incorrectly* as spam with probability 0.005.

You received one email, today. What is the probability that it really is spam given that it is flagged as spam?

Define the events

$A =$

and

$B =$

We know

$\mathbb{P}(B) =$,

$\mathbb{P}(B^c) =$,

$\mathbb{P}(A \mid B) =$, and

$\mathbb{P}(A \mid B^c) =$.

so Bayes' Rule gives

$\mathbb{P}(B \mid A) =$

$=$

$\approx$.

Independent events

We say A and B are *independent* if the knowledge that A has occurred does not change the *probability* that B occurs. Mathematically, this is written

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)\,.$$

Since $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ an alternative definition is:

This definition covers the case $B = \emptyset$, which is always independent of every event.

Mutually Independent Events

We can extend this to *arbitrary* many events:

The events $A_1, A_2, \ldots,$ are said to be *(*mutually) independent if for any n and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})\,.$$

Example: Suppose that a couple have four children, where each child is either a boy or a girl, each with probability 1/2, independent of the outcome for any other child. What is the probability of the children being born in the order (boy, girl, girl, girl)?

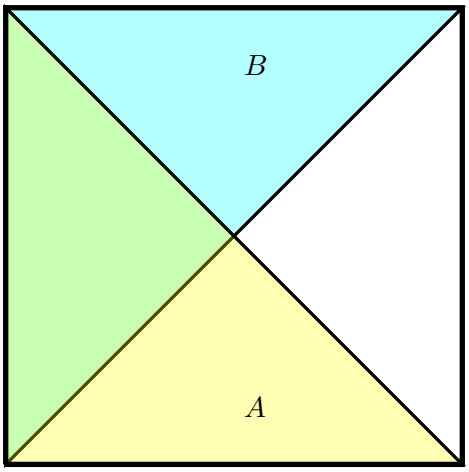
In our heads, we are thinking

$\mathbb{P}(BGGG) =$ $=$ $=$ $=$.

Mathematically, what's happening is we are treating the outcomes as events: Let B_1 be the event that the first child is a boy and let each G_i be the event that the ith child is a girl.

$\mathbb{P}(B_1G_2G_3G_4) =$ $=$ $=$.

Exercise. We uniformly select a point in the unit square. Show that the events $\{(x, y) : x + y \leqslant 1\}$ and $\{(x, y) : x - y \leqslant 0\}$ are independent.



$$\Omega = [0, 1] \times [0, 1]$$

Figure 3.2: $A = \{x + y \leqslant 1\}$ and $B = \{x - y \leqslant 0\}$.

We need to check if .

$$\mathbb{P}(A \cap B) = \quad ,$$

$$\mathbb{P}(A) = \quad , \text{ and}$$

$$\mathbb{P}(B) =$$

$$\Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Therefore A and B are events.

Remark. In most cases independence of events is a *model assumption*. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

Figure 3.3: The system fails if either component fails.

Example: Consider the system of Figure 3.3, with two components (1 and 2). The system fails if one (or both) of the components fails. Each component fails independently of the other with probability 0.01. What is the probability of a system failure?

Let A_i be the event that Component i fails. Let A be the event that the system fails.
Note that $A =$.
To apply independence, we need an intersection, so instead can look at $A^c =$.

$$\begin{aligned}
\mathbb{P}(A^c) &= \\
&= \\
\mathbb{P}(A) &= \\
&= \\
&\approx \quad .
\end{aligned}$$

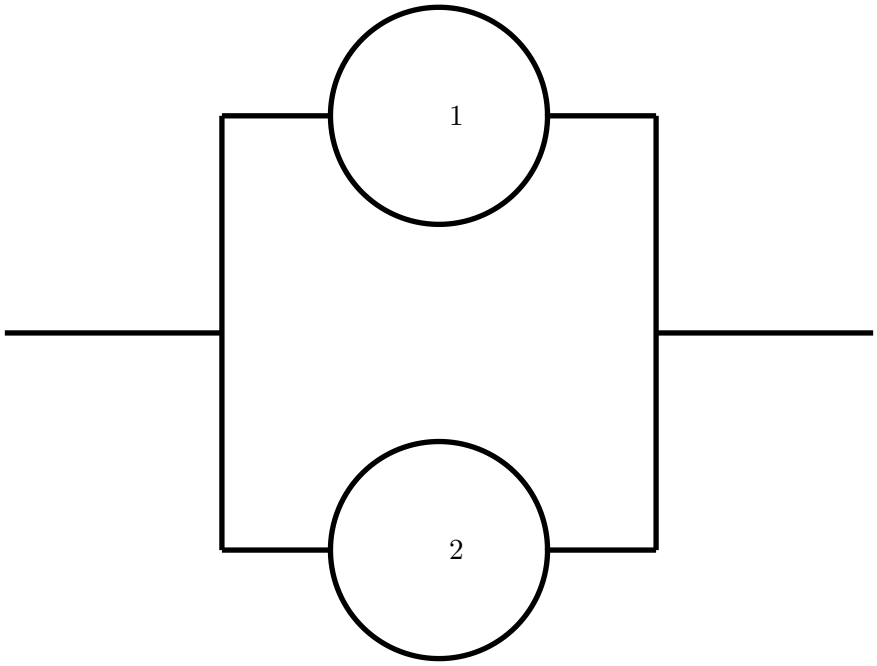


Figure 3.4: The system fails if both components fail.

Example: Consider the system of Figure 3.4, with two components (1 and 2), but this time, the system fails only if *both* of the components fails. Again, each component fails independently of the other with probability 0.01. What is the probability of a system failure?

Let A_i be the event that Component i fails. Let A be the event that the system fails.
Note that $A =$, this time.

$$\begin{aligned}
\mathbb{P}(A) &= \\
&= \\
&= \\
&= \quad .
\end{aligned}$$

4

Discrete Random Variables

By the end of this chapter you should:

- Know what a random variable is.
- Identify some common distributions for discrete random variables.
- Compute the expectation and variance of a discrete random variable.
- Be able to determine if two discrete random variables are independent from their joint distribution.
- Be able to manipulate joint, marginal and conditional pmf's.
- Compute the expectation and variance of a sum of discrete random variables.
- To use moment generating functions to calculate moments and identify the distribution of a random variable.

Random Variables

Many of the situations we have been concerned with have naturally had a number associated with them. For example, the number resulting from the rolling of a fair die. In other situations the connection between the sample space and a real number may not be so obvious, but is usually very useful.

Suppose that the sample space consists of 20 students chosen at random from this class. What are some numbers associated with a particular outcome ω?

$X(\omega) =$
$Y(\omega) =$
$Z(\omega) =$

The outcome of a random experiment is often expressed as a *number* or *measurement*.

Definition. A function X assigning a real number to every outcome $\omega \in \Omega$ is called a *random variable.*

Notation: It is often highly useful to have a function which indicates simply whether (or not) an item belongs to a particular set. Recalling that *events* (e.g. A) are sets, an indicator function is a *random variable* $\mathrm{I}_A : \Omega \to \{0, 1\}$ that takes on the value 1 if the outcome ω of our *random experiment* is in the set A (i.e. $\omega \in A$).

Example. We toss a coin three times, with the tosses being independent. The sample space is

$$\Omega =$$

i.e. sequences of length three of Hs (failures) and Ts (successes).

Consider the function $X : \Omega \to \{0, \ldots, 3\}$ which maps $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)$ to

$$X(\boldsymbol{\omega}) := \mathrm{I}_{\{\omega:\, \omega_1 = T\}}(\omega) + \mathrm{I}_{\{\omega:\, \omega_2 = T\}}(\omega) + \mathrm{I}_{\{\omega:\, \omega_3 = T\}}(\omega) \ .$$

X is a random variable. The set $\{X = k\}$ corresponds to the set of outcomes with exactly k successes. Hence, we can interpret X as the *total number of successes in three identical coin tosses.*

Let $p \in [0, 1]$ be the **probability of success** in a single coin toss. For example,

$$\begin{aligned} \mathbb{P}(\omega_1 = T,\, \omega_2 = T,\, \omega_3 = H) &= \\ &= \\ &= \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(\{\boldsymbol{\omega} \in \Omega : X(\boldsymbol{\omega}) = 2\}) \quad &= \\ &= \\ &= \\ &= \\ &= \qquad\qquad . \end{aligned}$$

Important Remarks

- Random variables are usually the *most convenient way to describe random ex-*

periments; they allow us to use intuitive notations for certain events, such as $\{X > 1000\}$, $\{\max(X, Y) \leqslant Z\}$, etc.

- Although mathematically a random variable is neither random nor a variable (it is a function), in practice we may interpret a random variable as the *measurement on a random experiment* which we will carry out *"tomorrow"*. However, all the *thinking about the experiment is done "today"*.
- We denote random variables by *upper case Roman letters*, X, Y, ...
- Numbers we get when we make the measurement (the outcomes of the random variables) are denoted by the *lower case* letter, such as x_1, x_2, x_3 for three values for X.

Example: Let X be the face value of a fair die, where $\Omega =$ and $X(\omega) = \omega$ (this is a simple random variable because the sample space already consists of real numbers).

Notation: Events such as $\{\omega \in \Omega : X(\omega) \leqslant x\}$ and $\{\omega \in \Omega : X(\omega) = x\}$, for some real number x, are abbreviated to $\{X \leqslant x\}$ and $\{X = x\}$. The corresponding probabilities of these events, $\mathbb{P}(\{X \leqslant x\})$ and $\mathbb{P}(\{X = x\})$, are abbreviated further to $\mathbb{P}(X \leqslant x)$ and $\mathbb{P}(X = x)$, respectively.

Let A be the event that the face value is even, so $A =$. Then

$$\begin{aligned} \mathbb{P}(A) &= \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

Types of Random Variable

Loosely speaking, the *set of all possible values* a random variable X that occur with positive probability is called the **support** of X, often denoted by $\text{supp}(X)$ or Ω_X.

- **Discrete** random variables can only take *isolated* values.

  For example: a count can only take non-negative integer values.
- **Continuous** random variables can take values in an *interval*.

  For example: rainfall measurements, lifetimes of components, lengths, ... are (at least in principle) continuous.

If we know, or can find the probabilities for all events defined by a random variable X, we know the **(probability) distribution** of X.

Question. How do we write down the distribution of a random variable?

Cumulative Distribution Function

The following function is defined for both continuous and discrete random variables.

Definition. The **cumulative distribution function** (cdf) of X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \mathbb{P}(X \leqslant x) \ .$$

Example: If X is the face value of a fair, six-sided die, then

$$\mathbb{P}(X \leqslant x) =$$

The cdf is shown in Figure 4.1.

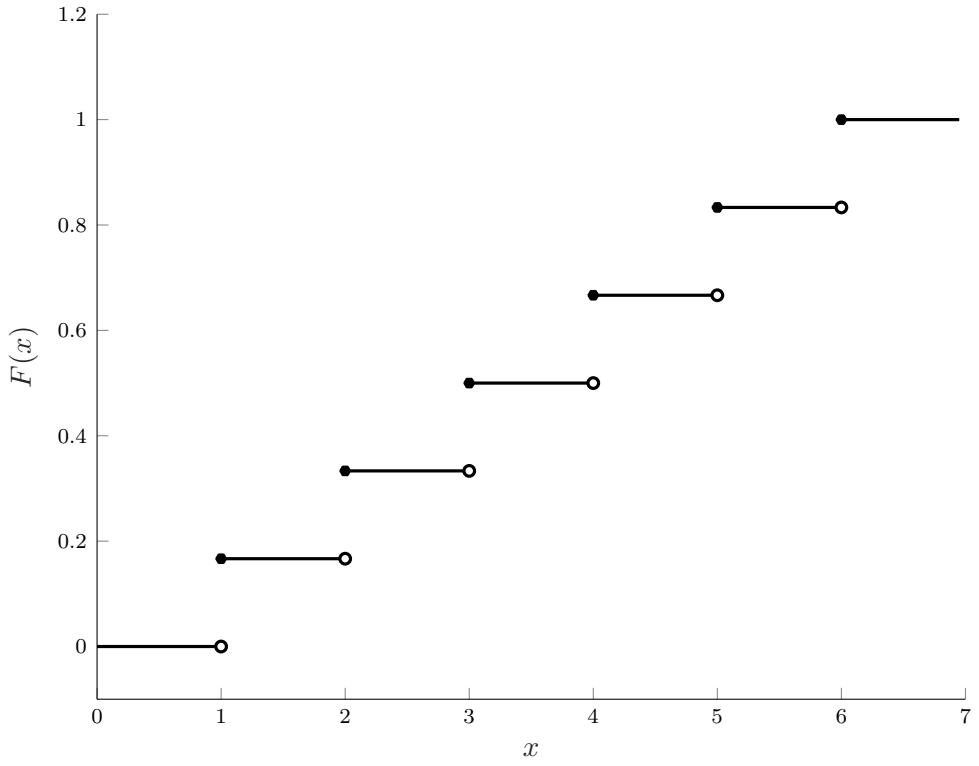


Figure 4.1: The cdf of the face value of a fair, six-sided die.

CDF Properties:

- $0 \leqslant F(x) \leqslant 1$.
- $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.
- F is *non-decreasing*: If $x < y$, then $F(x) \leqslant F(y)$.
- F is *right-continuous*: If $x_n \downarrow x$, then $\lim_{n\to\infty} F(x_n) = F(x)$.

For a *discrete* random variable X the cdf F is a **step function** with jumps of size $\mathbb{P}(X = x)$ at all the points $x \in \Omega_X$.

Probability Mass Function

Definition. For a *discrete* random variable X, the function $x \mapsto \mathbb{P}(X = x)$ is called the **probability mass function** (pmf) of X.

For a set B we have

$$f_X(B) = \mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x) \, .$$

Example: Roll a die and let X be its face value. X is discrete with support $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. If the die is fair, the probability mass function is given by

x	1	2	3	4	5	6	$\sum$
$f_X(x)$							

Note that $\mathbb{P}(X \in \Omega_X) =$.

Example: Roll two dice and let M be the largest face value showing. The distribution of M can be found to be

m	1	2	3	4	5	6	$\sum$
$f_M(m)$			$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	

or, as a formula:

$$f_M(m) = \mathbb{P}(M = m) = \frac{2m - 1}{36}, \text{ for } m = 1, 2, \ldots, 6 \, .$$

We can now work out the probability of *any* event defined by M so we know the distribution of M.

For example:

$\mathbb{P}(M > 4) =$

Common discrete distributions

Discrete Uniform Distribution

In the above example of rolling a die and recording the face value, all outcomes were equally likely. This is a uniform distribution on $\{1, 2, \ldots, 6\}$. In general, we say that a random variable X has a **discrete uniform distribution** on a finite set A if

$$\mathbb{P}(X = x) = \frac{1}{|A|}, \quad x \in A \, .$$

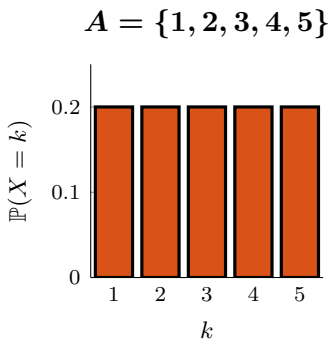


Figure 4.2: Probability mass function for $X \sim \mathsf{Uniform}(\{1, 2, 3, 4, 5\})$.

We write $X \sim \mathsf{DU}(A)$ or simply $X \sim \mathsf{Uniform}(A)$.

Examples.

-
-
-
-

Bernoulli Distribution

We say that a random variable X has a **Bernoulli distribution** with success parameter p if $\Omega_X = \{0, 1\}$, and

$$\mathbb{P}(X = 1) = p \qquad \text{and} \qquad \mathbb{P}(X = 0) = 1 - p.$$

We write $X \sim \mathsf{Bernoulli}(p)$.

A Bernoulli random variable describes the outcome of a *Bernoulli trial*.

Example: Flip a biased coin with probability heads p. The sample space Ω is {Heads, Tails}, we can define a random variable as:

ω	$X(\omega)$

This is a $\mathsf{Bernoulli}(p)$ random variable. If $p = 1/2$ it is *also* a discrete uniform random variable.

We can also define the random variable:

ω	$Y(\omega)$
Heads	0
Tails	1

This is also a $\mathsf{Bernoulli}(p)$ random variable.

A Bernoulli trial is a fancy way of talking about an experiment that either succeeds or fails, with only those outcomes being possible.

Suppose we define a new random variable as:

ω	$Z(\omega)$
Heads	10
Tails	0

This is NOT a $\mathsf{Bernoulli}(p)$ random variable. However $Z = 10\,X$ so it is a *function* of a Bernoulli random variable. Notice that we can easily find the pmf of Z from the pmf of X.

Binomial Bistribution

We say that a random variable X has a **Binomial distribution** with parameters n and p if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\ .$$

We write $X \sim \mathsf{Binomial}(n, p)$. We have encountered this distribution several times already.

A Binomial random variable is used to describe the *total number of successes* in a sequence of n independent Bernoulli trials with success probability p.

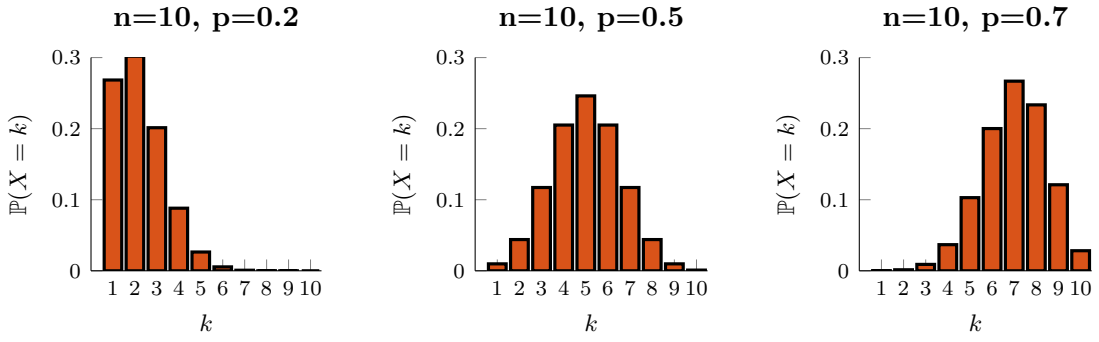


Figure 4.3: Probability mass function for $X \sim \mathsf{Binomial}(10, p)$ for $p = 0.2$, $p = 0.5$, and $p = 0.7$.

Example: A colection of 12 items contains 5 defectives. If you select one item from this collection (uniformly), the probability of the selected item being defective is $\boxed{}$.

A sample of size 4 is taken from this collection *with replacement*. Let X be the number of defectives in the sample. Then $X \sim \mathsf{Binomial}(\ ,\ \ )$.

$$\mathbb{P}(X = k) = \qquad\qquad , \quad k = 0, 1, \ldots, 4\,.$$

k	0	1	2	3	4	$\sum$
$\mathbb{P}(X = k)$	.116	.331	.354	.169	.030	1

Question. Can you give another example of a Binomial random variable?

We can write a function in MATLAB that generates a realisation from a $\mathsf{Binomial}(n, p)$ random variable using only the built in `rand` function as follows.

```
function output = Binomial(n,p)
    output = sum(rand(1,n) < p);
end
```

After saving this as 'Binomial.m' to our current working directory we can call this function as follows.

```
Binomial(100, 0.4)
ans =
    46

Binomial(100, 0.4)
ans =
    39
```

The function `binornd` built into MATLAB is a more sophisticated version of the function we just wrote. For more information on this function type "help binornd" into the MATLAB command line.

Geometric Distribution

Suppose that you have set up a printer on the local area network of your office. Suppose also that time is slotted into discrete units. In each time slot a single print job arrives with probability p indendently of the arrival of jobs at other time slots. We now have a sequence of independent Bernoulli random variables.

Question. What is the pmf of the time until the first job arrives?

$\mathbb{P}(X = 5) =$

$\mathbb{P}(X = k) =$

We say that the random variable X has a **Geometric distribution** with parameter p and we write $X \sim \mathsf{Geometric}(p)$.

Notice that after each arrival the time until the next arrival is again $\mathsf{Geometric}(p)$.

A Geometric random variable is used to describe the *time of first success* in a sequence of independent Bernoulli trials with success probability p.

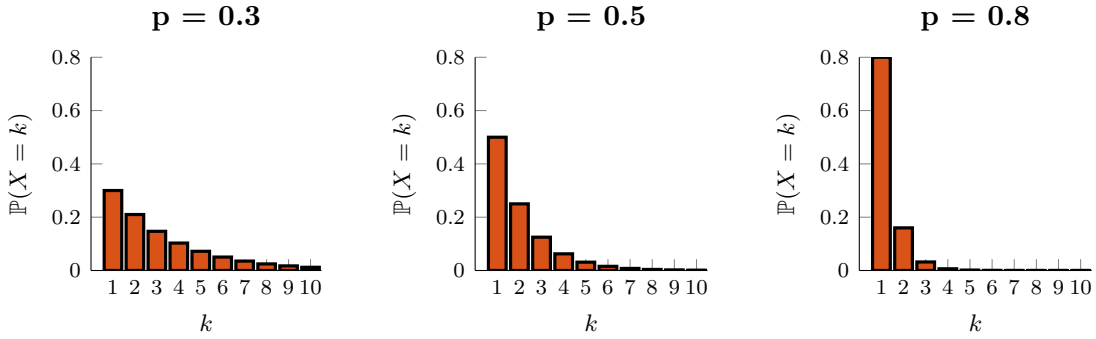


Figure 4.4: Probability mass function for $X \sim \mathsf{Geometric}(p)$ for $p = 0.3$, 0.5, and 0.8.

Question. Given an outcome $\omega \in \Omega$, what are the possible range of outcomes for $X(\omega) \sim \mathsf{Geometric}(p)$?

$$P(X = k) = (1-p)^{k-1}\, p\,, \quad k \in \{\qquad\qquad\}\,.$$

Remark. Be careful as other sources may define a Geometric random variable as the number of trials *before the first success*.

Example: Let X be the number of throws with a (fair) die, needed to get a six.

Then $X \sim \mathsf{Geometric}(\quad)$.

Examples.

-
-
-
-

Question. How would you write a function in MATLAB that generates a realisation from a $\mathsf{Geometric}(p)$ random variable using only the built in `rand` function?

Question. Returning to the printer example, suppose that no print jobs have arrived after k time slots, what is the probability that the next job arrives in the $m+k$ time slot ($m > 0$)?

$$\mathbb{P}(X = m + k \,|\, X > k) =$$

$$=$$

$$=$$

$$=$$

We just saw that

$$\mathbb{P}(X = m + k \,|\, X > k) = \mathbb{P}(X = m)\,.$$

This means that the distribution of the time we must wait for the next print job to arrive is independent of the amount of time we have already waited.

The Geometric distribution is the only discrete distribution with this extremely useful property.

Poisson Distribution

Consider a $\mathsf{Binomial}(10000, 0.005)$ random variable. This corresponds to a situation where we have 10000 independent Bernoulli trials with success probability 0.005.

$$\mathbb{P}(X = k) = \binom{10000}{k} 0.005^k \, 0.995^{10000-k}\,, \quad k = 0, 1, \ldots, 10000\,.$$

Question. Can any of you calculate $\mathbb{P}(X = 50)$ on your laptop? (Try to do this now!)

```
nchoosek(10000, 50)*0.005^50*0.995^9950
Warning: Result may not be exact. Coefficient is greater than
9.007199e+15 and is only accurate to 15 digits
> In nchoosek at 92
ans =
    0.0565
```

A Poisson distribution is the *limit of Binomial* distributions in the following sense:

Let $X_n \sim \mathsf{Binomial}(n, \lambda/n)$ with $\lambda > 0$ and X is such that

$$\mathbb{P}(X = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}\,, \quad k = 0, 1, 2 \ldots$$

then

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)\,,$$

for all k. We will justify this limiting result later when we look at probability generating functions.

We say that the random variable X has a **Poisson distribution** with parameter $\lambda > 0$. When $\lambda = 0$, the corresponding pmf is just

$$\mathbb{P}(X = 0) = 1, \quad \text{and} \quad \mathbb{P}(X = k) = 0 \quad \text{for } k \geqslant 1.$$

We write $X \sim \mathsf{Poisson}(\lambda)$.

Question. For $X \sim \mathsf{Poisson}(50)$ can any of you calculate $\mathbb{P}(X = 50)$ on your laptop? (Try to do this now!)

```
exp(-50)*50^50/factorial(50)
ans =
    0.0563
```

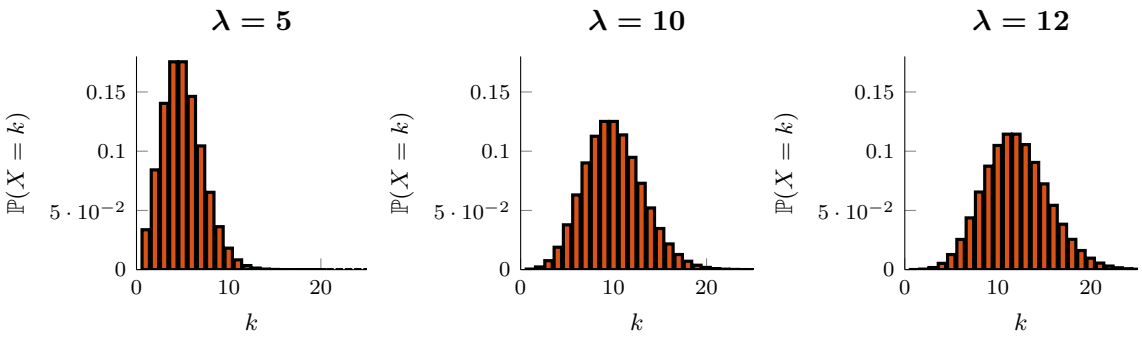


Figure 4.5: Probability mass function for $X \sim \mathsf{Poisson}(\lambda)$ for $\lambda = 5$, 10, and 12.

Note that unlike a Binomial random variable, which is restricted to take on values in $\{0, 1, \ldots, n\}$, a Poisson random variable can take on any non-negative integer value.

Examples.

-
-
-
-

We can generate a list of realisations from a $\mathsf{Poisson}(10)$ random variable in MATLAB as follows:

```
>> poissrnd(10,1,5)
ans =
    10     8    10     9    11
```

A Printer Example

Suppose that print jobs during the first hour of each day last week for the printer in Priestley arrived in slotted time according to Figure 4.

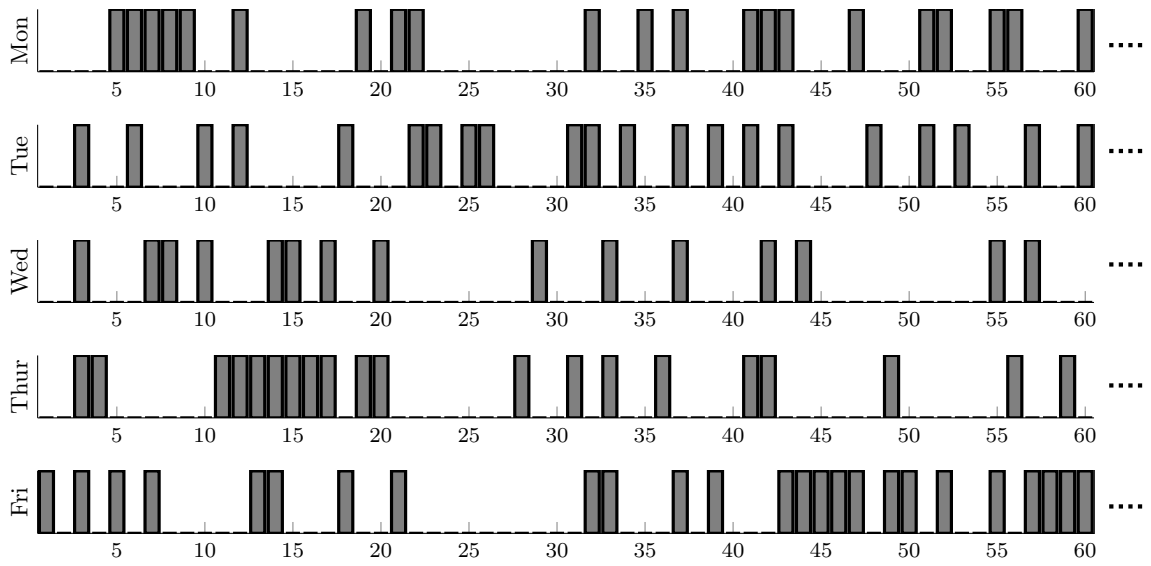


Figure 4.6: Print jobs in slotted time for first hour day during a week.

Question. How do we model the arrival of jobs in a given timeslot? (Assuming all time slots are independent.)

In this case the sample space for any given day contains all of the potential arrival sequences. For example

$$\omega = (0, 0, 1, 1, 0, \dots) .$$

Match the description on the left with a choice of states and then with a distribution.

Number of print jobs in any chosen n slots.	$\{1, 2, \dots\}$	Geometric$(1-p)$
Number of print jobs in any chosen n slots. (Large n and small p.)	$\{\dots, -2, -1, 0, 1, \dots\}$	Poisson(λ)
	$\{0, 1, \dots, n\}$	Bernoulli(p)
Indicator of print job arriving in a time slot.	$\{0, 1\}$	Geometric(p)
Time between print jobs.	$\{0, 1, 2, \dots\}$	Binomial(n, p)

Expected Value

Definition. Let X be a *discrete* random variable. The **expected value** (or **mean value** of X), denoted by $\mathbb{E}\,X$, is defined by

$$\mathbb{E}\,X = \sum_x x\,\mathbb{P}(X = x)\ .$$

This number, sometimes written as μ_X, is a measure of location for the distribution.

Example. Find $\mathbb{E}\,X$ where X is the outcome of a roll of a fair die.

Since $\mathbb{P}(X = 1) = \ldots = \mathbb{P}(X = 6) = 1/6$

$$\mathbb{E}\,[X] = \qquad .$$

Note: $\mathbb{E}\,[X]$ is not necessarily a possible outcome of the random experiment as in the previous example.

Example. Find $\mathbb{E}\,[X]$ for $X \sim \mathsf{Binomial}(n, p)$.

$$\begin{aligned}
\mathbb{E}\,[X] &= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&= \\
&=
\end{aligned}$$

Let X be a discrete random random variable and let g be a real-valued function defined on the support of X. We can define a new random variable Y as $Y := g(X)$. Note that Y is also a discrete random variable. The probability mass function of Y is

$$\mathbb{P}(Y = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x).$$

The expected value of Y is then

$$\mathbb{E}\,Y = \sum_y y\mathbb{P}(Y=y) = \sum_y y \sum_{x:g(x)=y} \mathbb{P}(X=x) = \sum_x g(x)\mathbb{P}(X=x).$$

This leads to the following natural defintion for the expectation of $g(X)$ for any real valued function g.

Definition. If X is a *discrete* random variable, then for any real-valued function g

$$\mathbb{E}\,[g(X)] = \sum_x g(x)\,\mathbb{P}(X=x)\,.$$

Example. Find $\mathbb{E}\,[X/n]$ where $X \sim \mathsf{Binomial}(n,p)$. We have

$\mathbb{E}\,[X/n] =$

In general, for any random variable X and real constants a and b

$$\mathbb{E}\,[aX+b] = \qquad .$$

Example. Find $\mathbb{E}\,[X]$ for $X \sim \mathsf{Poisson}(\lambda)$.

$\mathbb{E}\,[X] =$

$=$

$=$

$=$

$=$

$=$

$=$

$=$

Variance

Definition. The variance of a random variable X, denoted by $\mathrm{Var}(X)$ is defined by

$$\mathrm{Var}(X) = \mathbb{E}\,(X - \mathbb{E}\,X)^2\,.$$

This *number*, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X. The square root of the variance is called the **standard deviation** and is denoted by σ_X.

It may be regarded as a measure of the *consistency* of outcome, as a smaller value of $\mathrm{Var}(X)$ implies that X is more often near $\mathbb{E}\,X$ than for a larger value of $\mathrm{Var}(X)$.

We finally list two properties of the variance:

- $\mathrm{Var}(X) = \mathbb{E}\,[X^2] - (\mathbb{E}\,X)^2$; and
- $\mathrm{Var}(a\,X + b) = a^2\,\mathrm{Var}(X)$.

Example. Find $\mathrm{Var}(X)$ where X is the outcome of a roll of a fair die.

$$\mathbb{E}\,[X^2] =$$

$$=$$

$$\mathrm{Var}(X) =$$

The expectation and variance of the common discrete distributions can be found in the formula sheet. Familiarise yourself with these expressions. Are you able to derive them?

Example. From the formula sheet, if $X \sim \mathsf{Binomial}(n, p)$, then $\mathrm{Var}(X) = np(1-p)$. Find $\mathrm{Var}(X/n)$.

$$\mathrm{Var}(X/n) =$$

Question: How can we interpret this result?

Multiple Random Variables

When looking into detailed examples, we came across situations involving **multiple random variables** where *dependence* is an inherent model characteristic.

Examples.

1. We randomly select a person from a large population of twitter users and record the number of people they follow X and the number of people who follow them Y.

2. The number of swaps X and number of comparisons Y performed by a sorting algorithm such as quicksort.

3. We randomly select 20 people from a large population and ask their age. Number the people from 1 to 20, and let $X_1, \ldots, X_{20}$ be the measurements.

How can we specify a model these experiments?

We cannot just specify the pmf of the individual random variables. We also need to specify the "*interaction*" between the random variables. E.g., in Example 2, if the sorting algorithm needs to perform a large number of swaps X, then it also probably needs to do a large number of comparisons Y.

We need to specify the **joint distribution** of all the random variables $X_1, \ldots, X_n$ involved in the experiment. In other words, we need to specify the distribution of the random **vector** $\boldsymbol{X} := (X_1, \ldots, X_n)$. The **joint cumulative distribution function** F is defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) \; .$$

This completely specifies the probability distribution of the vector $\boldsymbol{X}$. However, if the X_i's are discrete, it suffices to only know the **joint probability mass function**.

Definition. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a *discrete* random vector. The function

$$(x_1, \ldots, x_n) \mapsto \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

is called the **joint probability mass function** of $\boldsymbol{X}$.

We will often just work with pairs of random variables (X, Y) having a joint pmf $f_{X,Y}$. The generalisation to multiple random variables is usually straightforward.

Example: Suppose that a fair, six-sided die is rolled with an independently-tossed fair coin. Let X be the face value of the die, in $\{1, 2, 3, 4, 5, 6\}$, and let Y be the outcome of the coin, in $\{0, 1\}$.

	x						
y	1	2	3	4	5	6	$\sum$
0		$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$		$\frac{1}{2}$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Example. In a box are three dice. Die 1 is a normal die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces.

The experiment consists of selecting a die at random, followed by a roll of that die. Let X be the die number that is selected, and let Y be the face value of that die. The joint pmf of X and Y is specified below.

x \ y	1	2	3	4	5	6	$\sum$
1		$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$		$\frac{1}{3}$
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$			$\frac{1}{3}$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Marginal Distributions

Consider a set B that consists of pairs of points, where each pair is of the type (x, y), for real numbers x and y. For example,

$B =$

$B =$

This latter set describes the possible (x, y) values from the previous example.

We have for a set B containing elements of the type (x, y),

$$f_{X,Y}(B) = \mathbb{P}((X, Y) \in B) = \sum_{(x,y)\in B} \mathbb{P}(X = x, Y = y) \, .$$

The pmf's f_X of X, the so-called **marginal** pmf, can be found by *summing up* over respectively the y's:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y) \, .$$

Similarly, the marginal pmf of Y can be found by summing up over the x's.

Question: Is this another version of the Law of Total Probability?

The Law of Total Probability gives

$f_X(x) = \mathbb{P}(X = x) =$

Independence of Discrete Random Variables

An important way of **creating** joint pmf's is by starting with the marginal pmf's of X and Y and then to define the events $\{X = x\}$ and $\{Y = y\}$ to be *independent*, for all x and y.

We then have (from the definition of independent events)

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)\ .$$

or put another way

$f_{X,Y}(x, y) =$

We call the *random variables* X and Y independent when this holds.

There is an important difference between independent random variables and independent events. Recall that if x and y are given (fixed) then the *events* $\{X = x\}$ and $\{Y = x\}$ being independent only means

$f_{X,Y}(x, y) =$

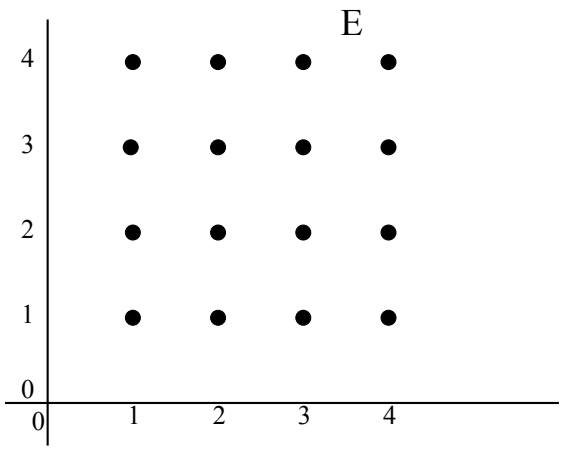
not for all x and y.

Note the similarity of these operations to those we performed earlier for events.

Example. Repeat the previous experiment with three ordinary dice. Since the events $\{X = x\}$ and $\{Y = y\}$ should be independent, each entry in the pmf table is $\frac{1}{3} \times \frac{1}{6}$.

Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent (which are not?). Hence the random variables X and Y are not considered to be independent.

Example. We draw at random a point (X, Y) from the 16 points on the square E below.



Clearly X and Y are independent.

Expectation Revisited

Similar to the one-dimensional case, the expected value of $Z = g(X, Y)$ can be evaluated as

$$\mathbb{E} = \sum_x \sum_y g(x, y)\mathbb{P}(X = x, Y = y),$$

and in general, the expected value of $Z = g(X_1, \ldots, X_n)$ can be evaluated as

$$\mathbb{E} = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Example. We will often be interested in sums of random variables. Find the expected value of $X + Y$.

$$\begin{aligned} \mathbb{E}[X+Y] &= \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

In general, suppose that $X_1, \ldots, X_n$ are random variables measured on the same random experiment. For arbitrary constants $b_0, b_1, \ldots, b_n$, we have

$$\mathbb{E}[b_0 + b_1 X_1 + \cdots b_n X_n] =$$

Note:

Example. Suppose that $X_1, \ldots, X_n$ are independent $\mathsf{Bernoulli}(p)$ random variables. Compute $\mathbb{E}[X_1 + \cdots + X_n]$.

Recall that if $X_i \sim \mathsf{Bernoulli}(p)$, then $\mathbb{P}(X_i = x) = p^x (1-p)^{1-x}$ for $x = 0, 1$.

$$\begin{aligned} \mathbb{E} X_i &= \\ &= \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[X_1 + \cdots + X_n] &= \\ &= \end{aligned}$$

We previously computed the expectation of a random variable having a $\mathsf{Binomial}(n, p)$ distribution. This required considerable effort. We know that a $\mathsf{Binomial}(n, p)$ random

variable describes the total number of sucesses in a sequence of n Bernoulli trials with success probability p. In other words, if $X_1, \ldots, X_n$ are independent $\mathsf{Bernoulli}(p)$ random variables, then $\sum_{i=1}^{n} X_i \sim \mathsf{Binomial}(n, p)$. Using linearity of expectations has greatly simplified this calculation.

Example. Suppose that X and Y are two *independent* random variables measured on the same random experiment. Find the expected value of XY.

$$\mathbb{E}[XY] =$$

$$=$$

$$=$$

$$=$$

Important: Suppose that $X_1, \ldots, X_n$ are *independent* random variables measured on the same random experiment. We have

$$\mathbb{E}[X_1 X_2 \cdots X_n] =$$

Variance and Covariance

In the previous subsection, we saw that the expected value of the sum of random variables is equal to the sum of the expectations of the individual random variables. This gives a measure of location for the distribution of a sum of random variables. We have also seen that the spread of a distribution can be described by the variance. It is natural, therefore, to consider the variance of a sum of random variables. In general, the variance of a sum of random variables does not equal the sum of the variances, but includes an extra term called the covariance.

Definition. The covariance of a two random variables X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)] .$$

It is a measure of the amount of linear dependence between the two random variables.

Example. Suppose that X and Y are two random variables measured on the same random experiment. Find the variance of $X + Y$.

$$\begin{aligned}\text{Var}[X+Y] &= \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

Recall that if X and Y are two independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. It follows that if X and Y are two independent random variables, then

$$\text{Cov}(X,Y) =$$

and

$$\text{Var}(X+Y) =$$

Important: Suppose that $X_1, \ldots, X_n$ are *independent* random variables measured on the same random experiment. We have

$$\text{Var}(X_1 + X_2 + \cdots + X_n) =$$

Example. Find the variance of $X \sim \mathsf{Binomial}(n,p)$.

Let $X_1, \ldots, X_n$ be independent $\mathsf{Bernoulli}(p)$ random variables. Then $\sum_{i=1}^n X_i$ has a $\mathsf{Binomial}(n,p)$ distribution. We can therefore compute the variance of X by computing the variance of $\sum_{i=1}^n X_i$.

Recall that if $X_i \sim \mathsf{Bernoulli}(p)$, then $\mathbb{P}(X_i = x) = p^x(1-p)^{1-x}$ for $x = 0, 1$.

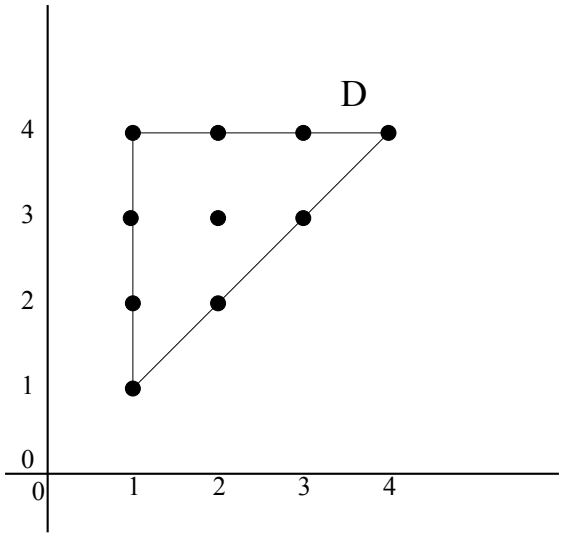
$$\begin{aligned}\text{Var}(X_i) &= \\ &= \end{aligned}$$

Therefore,

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) =$$
$$=$$

Conditional probability mass function

Example. We draw at random a point (X, Y) from the 10 points on the triangle D below.



The joint and marginal pmf's are easy to determine:

$$\mathbb{P}(X = i, Y = j) = \qquad (i, j) \in D\,,$$
$$\mathbb{P}(X = i) = \frac{5 - i}{10}\,, \qquad i \in \{1, 2, 3, 4\}\,,$$
$$\mathbb{P}(Y = j) =$$

Clearly X and Y are *not independent*. In fact, if we know that $X = 2$, then Y can only take the values $j = 2, 3$ or 4.

The corresponding probabilities are

$$f_{Y\,|\,X}(j, 2) = \begin{cases} \mathbb{P}(Y = j \,|\, X = 2) = \dfrac{\mathbb{P}(Y = j, X = 2)}{\mathbb{P}(X = 2)} = \dfrac{1/10}{3/10} = \dfrac{1}{3} & \text{if } j \in \{2, 3, 4\}, \\ 0 & \text{otherwise.} \end{cases}$$

We thus have determined the **conditional pmf** of Y given $X = 2$.

Definition. If X and Y are *discrete* and $\mathbb{P}(X = x) > 0$, then the probabilities

$$f_{Y\,|\,X}(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{f_{X,Y}(x, y)}{f_X(x)}\,,$$

for all y, give the **conditional pmf** of Y given $X = x$.

For a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ we also have the **chain rule**

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2 \mid X_1}(x_2 \mid x_1) \ldots f_{X_n \mid X_1,\ldots,X_{n-1}}(x_n \mid x_1, \ldots, x_{n-1})\,.$$

This is also known as *factorising* the joint distribution.

In the case $\mathbf{X} = (X_1, X_2)$ we have

$$f_{X_1\,,\,X_2}(x_1, x_2) =$$

The choice of how to factorise the distribution often depends on what we are modelling or what information is available. Another possible factorisation is given by

$$f_{X_1\,,\,X_2}(x_1, x_2) = f_{X_2}(x_2)$$

When X and Y are independent, this also gives us

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} =$$

Example. Let $Y \sim \mathsf{Poisson}(\lambda)$. The conditional on Y, X has a $\mathsf{Binomial}(Y, p)$ distribution. Find the joint pmf of (X, Y) and the marginal pmf of X.

To find the joint pmf of (X, Y):

$$f_{XY}(x, y) = f_{X \mid Y}(x \mid y) f_Y(y) =$$

To find the marginal pmf of X:

$$\begin{aligned} f_X(x) &= \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

Conditional Expectation

We have previously seen the conditional probability mass function and expectation. We now consider taking expectation with respect to a conditional probability mass function. Two important reasons for considering this are:

- Conditioning arguments can facilitate the computation of expectations.
- The conditional expectation can be viewed as a prediction of a random variable given certain available information.

There are two notions of conditional expectation. The first we will consider is the conditional expectation given an event.

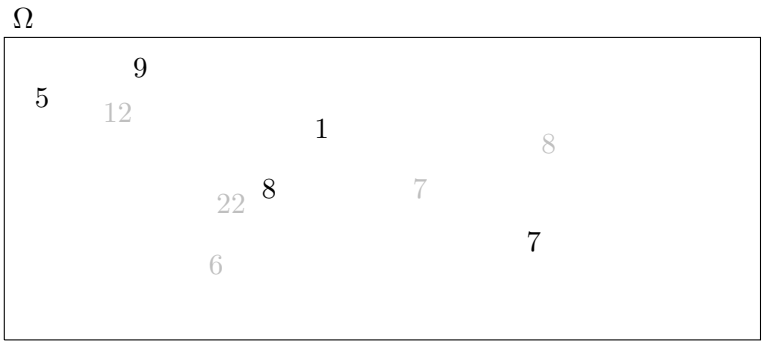
Definition. (Conditional expectation given an event) Let A be an event that occurs with positive probability. The conditional expectation of Y given A is

$$\mathbb{E}\,[Y \,|\, A] = \sum_y y\, \mathbb{P}(Y = y \,|\, A).$$

In particular, if $A = \{X = x\}$, then

$$\mathbb{E}\,[Y \,|\, X = x] = \sum_y y\, \mathbb{P}(Y = y \,|\, X = x) = \sum_y y\, f_{Y\,|\,X}(y \,|\, x).$$

Example. The sample space depicted below consists of black and grey numbers.



Assume that we will choose a colour number pair $\omega \in \Omega$ from this collection uniformly at random. Define the random variable $Y(\omega)$ as the number associated with the selection ω and the random variable $X(\omega)$ as

$$X(\omega) = \begin{cases} 1\,, & \text{if } \omega \text{ black,} \\ 0\,, & \text{if } \omega \text{ grey.} \end{cases}$$

Question. What is the expected value of Y if we know that X is equal to 1?

$$\mathbb{E}\,[Y \,|\, X = 1] =$$

Question. What is the expected value of our selection if we know that X is equal to 0?

$$\mathbb{E}[Y \mid X = 0] =$$

We can relate the conditional expectation $\mathbb{E}[Y \mid X = x]$ to the unconditional expectation $\mathbb{E}[Y]$ as follows: For any random variables X and Y defined in the same random experiement

$$\begin{aligned}\sum_x \mathbb{E}[Y \mid X = x]\,\mathbb{P}(X = x) &= \\ &= \\ &= \\ &= \\ &=\end{aligned}$$

Example. Compute $\mathbb{E}[Y]$ from $\mathbb{E}[Y \mid X = 1]$ and $\mathbb{E}[Y \mid X = 0]$.

The other notion of conditional expectation is the conditional expectation given a random variable.

Definition. (Conditional expectation given a random variable) The expression $\mathbb{E}[Y \mid X]$ is a random variable $g(X)$ that takes the value $\mathbb{E}[Y \mid X = x]$ when $X = x$.

Question. What is the support of the random variable $\mathbb{E}[Y \mid X]$?

Since $\mathbb{E}[Y \mid X]$ is a random variable, we may take its expectation $\mathbb{E}[\mathbb{E}[Y \mid X]]$.

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}[Y \mid X]\right] &:= \sum_x \mathbb{E}[Y \mid X = x]\,\mathbb{P}(X = x) \\ &= \mathbb{E}[Y].\end{aligned}$$

This is very useful when we wish to know the expectation of a random sum of random variables.

Example. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with common mean μ, and let

$$S_n = \sum_{i=1}^{n} X_i \,.$$

$$\mathbb{E}\,[S_n] =$$

Now let N be a random variable which takes values in the non-negative integers. Then,

$$\begin{aligned} \mathbb{E}\,[S_N] &= \\ &= \\ &= \end{aligned}$$

Two final properties of conditional expectation that we may have occassion to use are:

Property. Let $Y_1, Y_2, \ldots, Y_n$ and X be random variables defined in the same random experiment. Then

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i \Big| X\right] = \sum_{i=1}^{n} \mathbb{E}\,[Y_i \,|\, X].$$

Property. If X and Y are independent, then $\mathbb{E}\,[Y \,|\, X]$ is constant and equal to $\mathbb{E}\,[Y]$.

$$\begin{aligned} \mathbb{E}\,[Y \,|\, X = x] &= \\ &= \end{aligned}$$

Moment Generating Functions

Let X be a *non-negative* and *integer-valued* random variable.

The **moment generating function** (MGF) of X is the function $M : \mathbb{R} \to (0, \infty)$ defined by

$$M_X(s) := \mathbb{E}\,\mathrm{e}^{sX} = \sum_{n=0}^{\infty} \mathrm{e}^{sn} \mathbb{P}(X = n)\,,$$

defined for all $s \in \mathbb{R}$ for which $\mathbb{E}\,\mathrm{e}^{sX}$ exists (is finite). We will require that $M_X(s)$ exist for all s in some open interval containing the origin.

Question. What does $M_X(0)$ equal?

Example. Find the MGF of $X \sim \mathsf{Geometric}(p)$.

$$\begin{aligned} G_X(s) &= \\ &= \\ &= \\ &= \qquad . \end{aligned}$$

Whenever $M_X(s)$ is defined, it can be determined from the distribution of X. Is the converse true? Given a moment generating function $M_X(s)$, can the distribution of X be determined? Fortunately, the answer is YES; moment generating functions have the **uniqueness property** – *two pmf's are the same if and only if their MGF's are the same.*

Recall from the properties of expectation that when X and Y are independent

$$\mathbb{E}\left[g(X)\, h(Y)\right] = \mathbb{E}\, g(X) \mathbb{E}\, h(Y)\,.$$

For the MGF we just defined this implies, for $Z = X + Y$, with X and Y independent (non-neg., integer-valued) random variables:

$$M_Z(s) = M_{X+Y}(s) =$$

Exercise: Prove the formula above.

$$\begin{aligned} M_Z(s) &= \\ &= \\ &= \\ &= \qquad\qquad \text{(by independence)} \\ &= \end{aligned}$$

Using this fact and the uniqueness property, to find the distribution of a sum of a collection of random variables we simply need to multiply the MGF's of the random variables.

Example. Find the MGF of $X \sim \mathsf{Binomial}(n, p)$. Recall that if $X_1, \ldots, X_n$ be independent $\mathsf{Bernoulli}(p)$ random variables. Then $\sum_{i=1}^{n} X_i$ has a $\mathsf{Binomial}(n, p)$ distribution. We can therefore compute the MGF of X by computing the MGF of $\sum_{i=1}^{n} X_i$.

Recall that if $X_i \sim \mathsf{Bernoulli}(p)$, then $\mathbb{P}(X_i = x) = p^x(1-p)^{1-x}$ for $x = 0, 1$.

$$M_{X_i}(s) =$$

$$=$$

Therefore,

$$M_X(s) =$$

$$=$$

Example. In the summary of formulas at the start of the workbook, the PGF of $X \sim \mathsf{Poisson}(\lambda)$ is given as:

$$M_X(s) =$$

Hence, for $X \sim \mathsf{Poisson}(\lambda)$ and $Y \sim \mathsf{Poisson}(\mu)$ (indep.), the MGF of $Z = X + Y$ is:

$$M_Z(s) =$$

Which shows that $Z \sim$.

A useful property of the MGF is that we can obtain the *moments* of X by *differentiating* M and evaluating it at $s = 0$.

Differentiating $M(s)$ with respect to s gives

$$\begin{aligned} M'(s) &= \frac{\mathrm{d}\mathbb{E}\,\mathrm{e}^{sX}}{\mathrm{d}s} = \mathbb{E}\,X\mathrm{e}^{sX} \\ M''(s) &= \frac{\mathrm{d}\mathbb{E}\,X\mathrm{e}^{sX}}{\mathrm{d}s} = \mathbb{E}\,X^2\mathrm{e}^{sX} \\ M'''(s) &= \mathbb{E}\,X^3\mathrm{e}^{sX} \\ &\vdots \\ M^{(k)}(0) &= \mathbb{E}\left[X^k\right] \end{aligned}$$

In particular

$$\mathbb{E}[X] =$$

and

$$\text{Var}(X) =$$

Exercise: Prove this variance formula.

$$\text{Var}(X) =$$

but

$$M'_X(0) =$$
$$M''_X(0) =$$
$$=$$
$$\Rightarrow \mathbb{E}\left[X^2\right] =$$

Exercise. Using the MGF find the variance of a Geometric(p) distribution.

$M'_X(s) =$

$M''_X(s) =$

so

$$\text{Var}(X) =$$
$$=$$
$$=$$

Finally, we stated earlier that the Poisson distribution can be viewed as the limit of a sequence of Binomial distributions with the number of trials increasing and the success probability decreasing at a particular rate. We can justify this using MGFs.

Let $X_n,\ n = 1, 2, \ldots$ be a sequence of non-negative integer valued random variables with MGFs $M_{X_n}(s)$ and let $M_X(s)$ be the MGF of X. If $M_{X_n}(s) \to M_X(s)$ for all s in some neighbourhood of 0, then

$$\lim_n \mathbb{P}(X_n = k) = \mathbb{P}(X = k), \quad \text{for all } k = 0, 1, 2 \ldots$$

We say that X_n *converges in distribution* to X.

Example. Consider the sequence of random variable $X_n \sim \mathsf{Binomial}(\mathsf{n}, \lambda/\mathsf{n})$. We can show that X_n converges in distribution to a $\mathsf{Poisson}(\lambda)$ random variable.

$$M_{X_n}(s) =$$

It is known that $\lim_{n\to\infty}(1 + x/n)^n = e^x$. So

$$\lim_{n\to\infty} M_{X_n}(s) =$$

which we can identify as the MGF of a Poisson distribution with mean λ.

Application: Run time of quicksort

Consider a list of n distinct numbers which we want to sort into increasing order. The quicksort algorithm begins by choosing an element from the list called the pivot. The pivot is compared with all other elements in the list and the list is then divided into two sublists:

- one comprising those elements less than the pivot,
- the other comprising those elements greater than the pivot.

This procedure is then repeated on the sublists until the entire list is sorted.

(Initial list)	10	5	3	7	9	2	1
(With 7 as pivot)	5	3	2	1	**7**	10	9
(With 3 as pivot for left sub-list)	2	1	**3**	5	7	10	9
...							
(List sorted)	1	2	3	5	7	9	10

If the pivot is simply taken to be the first element of the list and the list is already sorted then all pairs of numbers will need to be compared. Therefore, the number of comparisons is $\boxed{}$.

On the other hand, if the pivot is selected uniformly at random from the list, then the number of comparisons tends to be much smaller.

Let X_n be the random variable giving the number of comparison required to sort a list of n items. By convention $\mathbb{P}(X_0 = 0) = 1$. To compute $\mathbb{E}\, X_n$ we will condition of the selection of the first pivot

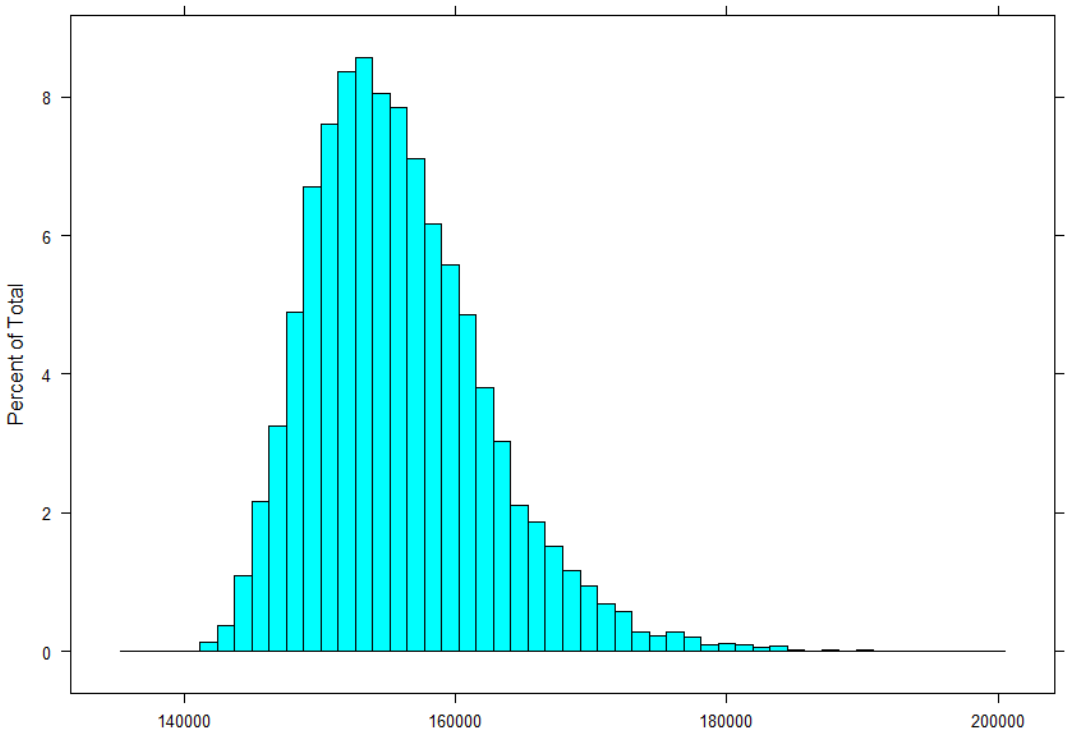


Figure 4.7: Histogram of the number of comparisons made by quicksort when sorting random permutations of $\{1, 2, \ldots, 10000\}$.

$$\mathbb{E}\, X = \mathbb{E}\,[\mathbb{E}\,[X_n \,|\, \text{first pivot}]]$$

$$=$$

$$=$$

$$=$$

Writing e_n for $\mathbb{E}\, X_n$ gives the recurrence equation

$$e_n = (n-1) + \frac{1}{n}\sum_{k=1}^{n}(e_{k-1} + e_{n-k})$$

$$=$$

To solve this recurrence we need to do some re-arranging:

$$
\begin{aligned}
e_n &= && \text{[multiply both sides by } n\text{]} \\
ne_n &= && \text{[subtract } (n-1)e_{n-1} \text{ from both sides]} \\
ne_n - (n-1)e_{n-1} &= \\
&= \\
ne_n &= \\
\frac{e_n}{n+1} &= && \text{[now iterating gives]} \\
\frac{e_n}{n+1} &= && \text{[use partial fractions]} \\
&= \\
&=
\end{aligned}
$$

We note that $\ln n \leqslant \sum_{k=1}^{n} \frac{1}{k} \leqslant \ln n + 1$. To conclude, we see that

$$\lim_{n\to\infty} \frac{e_n}{2n \ln n} = 1.$$

5

Continuous Random Variables

By the end of this chapter you should be able:

- To identify common continuous distributions.
- To compute the expectation and variance of a continuous random variable.
- To determine the probability density function of a transformed continuous random variable.
- To manipulate joint, marginal and conditional probability density functions.
- To compute the probability density function of a sum of two independent random variables using the convolution formula.
- To use moment generating functions to calculate moments and identify the distribution of a random variable.
- To understand the properties of the multivariate normal distribution.

So far we have dealt exclusively with discrete random variables. There are many situations where it may be more appropriate to use a continuous random variable. We will first recall the definitions of random variable and cumulative distribution function.

Definition. A function X assigning a real number to every outcome $\omega \in \Omega$ is called a **random variable**.

Definition. The **cumulative distribution function** (cdf) of X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \mathbb{P}(X \leqslant x) \, .$$

The following properties of the cumulative distribution function are basic consequences of the axioms of probability (Chapter 4):

- $0 \leqslant F(x) \leqslant 1$.
- $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

- F is *non-decreasing*: If $x < y$, then $F(x) \leqslant F(y)$.
- F is *right-continuous*: If $x_n \downarrow x$, then $\lim_{n\to\infty} F(x_n) = F(x)$.

We saw that for discrete random variables the cumulative distribution function is a step function. The cumulative distribution function of a continuous random variable has different behaviour.

Continuous Distributions

Definition. A random variable X is said to have a **continuous distribution** if its cumulative distribution function is continuous. A random variable with a continuous distribution is called a **continuous random variable**.

For a continuous random variable X and any $x \in \mathbb{R}$,

$$\mathbb{P}(X = x) =$$
$$=$$
$$=$$

Therefore, if X is a continuous random variable, then $\boxed{\mathbb{P}(X \leqslant x) =}$

Probability Density Function

For a continuous random variable X the probability $\mathbb{P}(X = x)$ is always 0. Hence, we cannot characterize the distribution of X via the probability mass function.

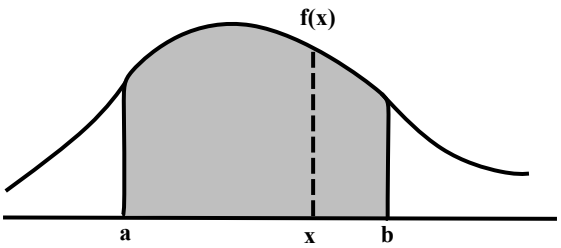
Instead, we have:

Definition. We say that a *continuous* random variable X has a **probability density function** (pdf) f_X if for all $x \in \mathbb{R}$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du\ .$$

Hence, for all a, b

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\, \mathrm{d}x.$$



By the Fundamental Theorem of Calculus if F_X is differentiable at x, then $\boxed{f_X(x) = \quad}$.

We can interpret the probability density function $f_X(x)$ as the *"infinitesimal"* probability that $X = x$. More precisely, for small $h > 0$,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_x^{x+h} f_X(u)\, du \approx h\, f_X(x)\ .$$

But, note carefully, $f_X(x)$ is not a probability. In particular, it is not true that $f_X(x) = \mathbb{P}(X = x)$, for all x and we may have $f_X(x) > 1$ for some values of x.

Basic properties of f_X:

- $f_X(x) \geqslant 0, \quad$ for all x;
- $\int_{-\infty}^{\infty} f_X(x)\, \mathrm{d}x = 1$

Any function f satisfying these properties is the pdf of some distribution.

Important: Although not all continuous distributions have a probability density function, we will not consider those distributions in this course. Henceforth, when we discuss continuous distributions, we will assume the existence of a pdf.

Example: Let X be the random variable having pdf

$$f_X(x) = \begin{cases} c(1 - x^2), & \text{if } x \in [-1, 1] \\ 0, & \text{if } x \notin [-1, 1]. \end{cases}$$

where c is a constant. What value must c take for f_X to be a valid pdf?

$$\int_{-\infty}^{\infty} f_X(x)\, dx =$$

Therefore, $c =$

What is the appropriate notion of expectation for continuous random variables?

Definition. Let X be a continuous random variable with pdf f_X. Then, the expected value of X is defined as:

$$\mathbb{E}\, X = \int_{-\infty}^{\infty} u f_X(u)\, du.$$

In a way that is analogous to the definition of expectation for discrete random variables, $\mathbb{E}\, X$ is a weighted average of the values in the support of X, weighted, that is, according to the density f_X. We can also take expectations of functions of random variables.

Definition. Let X be a continuous random variable with pdf f_X and let g be any real-valued function. Then, the expected value of $g(X)$ is defined as:

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(u) f_X(u)\, du.$$

When we come to look at transformations, we will see that these two definitions are consistent. All of the properties of expectation mentioned in the previous chapter hold

for continuous random variables. Furthermore, the definition and properties of the variance are the same. So to evaluate the variance of a random variable X, as before,

$$\mathrm{Var}(X) = \mathbb{E}\,[(X - \mathbb{E}\,X)^2] = \mathbb{E}\,[X^2] - (\mathbb{E}\,X)^2.$$

Example: Let X be the random variable having pdf

$$f_X(x) = \begin{cases} (1-x^2), & \text{if } x \in [-1,1] \\ 0, & \text{if } x \notin [-1,1]. \end{cases}.$$

What is the expected value and variance of X?

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

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

$$=$$

$$\mathrm{Var}(X) = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 =$$

Common continuous distributions

Uniform Distribution

Consider a random variable X taking values in the interval $[a, b]$. If all sub-intervals of equal length have the same probability of containing X, the we say X has a *continuous* **uniform distribution** on $[a, b]$.

Example: If $X \sim \mathsf{Uniform}[0, 1/2]$ then the pdf of X is

$$f_X(x) = \begin{cases} c, & x \in (0, 1/2), \\ 0, & x \notin (0, 1/2), \end{cases}$$

where c is some constant. What is c?

giving

What is $\mathbb{P}(X \leqslant 1/4)$?

$$\mathbb{P}(X \leqslant 1/4) =$$

In general, the probability density function of the uniform distribution on $[a, b]$ is

$$f_X(x) = \begin{cases} \quad , & x \in [a, b], \\ 0, & x \notin [a, b], \end{cases}$$

and the cumulative distribution function of the uniform distribution on $[a, b]$ is

$$F_X(x) = \begin{cases} 0, & x \in (-\infty, a] \\ \quad , & x \in [a, b], \\ 1, & x \in [b, \infty), \end{cases}$$

Since $\mathbb{P}(X = x) = 0$ for a continuous distribution, the uniform distributions on $[a, b]$, $[a, b)$, $(a, b]$ and (a, b) are all essentially the same.

The expected value and variance of the uniform distribution on $[a, b]$ are

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

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

$$=$$

$$\mathrm{Var}(X) = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2 =$$

$$=$$

Question. Let n be a positive integer. If X has a continuous uniform distribution on $(0, 1)$, what is the distribution of $Y = \lceil nX \rceil$? (*Note: $\lceil x \rceil$ means rounding x up to the nearest integer.*)

First note that Y is a discrete random variable taking values in $\{1, 2, \ldots, n\}$. We now want to determine the probability mass function of Y. For $y \in \{1, 2, \ldots, n\}$,

$$f_Y(y) = \mathbb{P}(Y = y) =$$
$$=$$
$$=$$
$$=$$
$$=$$

So Y has a [ ] on $\{1, 2, \ldots, n\}$.

Question. Suppose X has a continuous uniform distribution on (a, b) and let $c \in (a, b)$. What is the distribution of X conditioned on the event $\{X \leqslant c\}$?

For $x \in (a, c)$,

$$\mathbb{P}(X \leqslant x \,|\, X \leqslant c) =$$
$$=$$
$$=$$
$$=$$

So conditional on the event $\{X \leqslant c\}$, X has a [ ] on [ ].

Exponential Distribution

Example: Let X be the random variable describing the time until the first bug is reported in a computer program. Suppose the cdf of X is given by

$$F_X(x) = \mathbb{P}(X \leqslant x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-\lambda x}, & \text{if } x \geqslant 0, \end{cases}$$

where λ is a positive constant that may depend on the complexity of the program. What is the pdf of X?

If $x < 0$, then [$f_X(x) =$].

If $x > 0$, then [$f_X(x) =$].

However, $F_X'(0)$ does not exist. This doesn't matter; $f_X(0)$ can be set to anything non-negative, say $f_X(0) = \lambda$. In this case we get

$$f_X(x) =$$

Note that the "0 if $x < 0$" part of the function is essential.

We call the distribution with this pdf the **Exponential distribution** with rate parameter λ. We denote this distribution by $\mathsf{Exp}(\lambda)$. It is one of the most important distribution in *Applied Probability* due to its connection to continuous time Markov chains and the Poisson distribution (which we will see later) and its many useful properties. Like the discrete geometric distribution, the exponential distribution has the *memoryless* property.

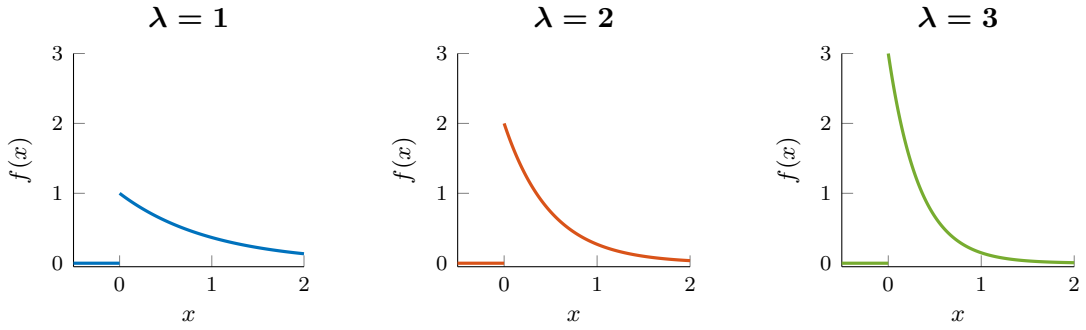


Figure 5.1: Probability density function for $X \sim \mathsf{Exp}(\lambda)$ for $\lambda = 1$, 2, and 3. Compare this to Figure 4.4.

Question. What is the probability that the time until the first bug is reported is greater than time x?

Assuming $x > 0$,

$$\begin{aligned} \mathbb{P}(X > x) &= \\ &= \\ &= \quad . \end{aligned}$$

The function $\mathbb{P}(X > x)$ is often called the "survivor function" or "reliability function" and is related to cumulative distribution function by $\mathbb{P}(X > x) = 1 - F_X(x)$.

Question. Given that no bug is found after y time units, what is the probability that it takes longer than time $x + y$ for the first bug to be found?

$$\begin{aligned} \mathbb{P}(X > x + y \,|\, X > y) &= \\ &= \\ &= \\ &= \end{aligned}$$

Question. If $X \sim \mathsf{Exp}(\lambda)$, what is the distribution of $Y = \lceil X \rceil$? The support of the distribution of Y is $\{1, 2, \ldots\}$. The probability mass function of Y is given by

$$\begin{aligned} f_Y(y) = \mathbb{P}(Y = y) &= \\ &= \\ &= \\ &= \\ &= \end{aligned}$$

where $y \in \{1, 2, \ldots\}$. We can recognise this as the probability mass function of a [ ] distribution.

What is the expected value and variance of the $\mathsf{Exp}(\lambda)$ distribution? It will be useful to recall that for two functions u and v we may use *integration by parts* as follows

$$\int u(x)v'(x)\mathrm{d}x = u(x)v(x) - \int v(x)u'(x)\mathrm{d}x\,.$$

So

$$\begin{aligned} \mathbb{E}[X] &= \\ &= \\ &= \\ &= \\ \mathbb{E}[X^2] &= \int_0^\infty x^2 f_X(x)\,\mathrm{d}x = \int_0^\infty x^2\,\lambda e^{-\lambda x}\,\mathrm{d}x \\ &= \\ &= \\ &= \\ \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}\,X)^2 = \end{aligned}$$

Standard Normal Distribution

The normal (or *Gaussian*) distribution plays a central role in probability and statistics. Many observed quantities appear to follow a normal distribution. Later we will see that the central limit theorem states that, under mild conditions, the average of independent

random variables follows approximately a normal distribution. We will begin our study of the normal distribution by considering an important special case.

Definition. A continuous random variable X is said to have **standard normal distribution** if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}.$$

We will denote this by $X \sim \mathsf{Normal}(0, 1)$.

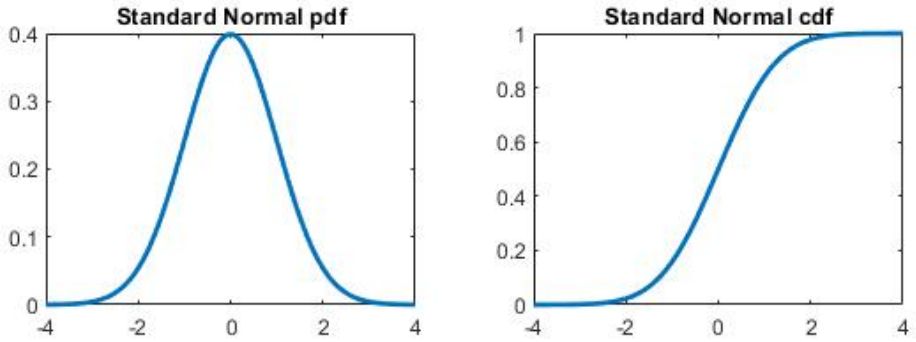


Figure 5.2: Probabilty density function and cumulative distribution function of the standard normal distribution.

There is no explicit expression for the cdf of the standard normal distribution. However, many software packages have methods for numerically evaluating the cdf, for example the `normcdf` function in MATLAB. In exams, we will use tables of the standard normal cdf like the one given at the start of these notes. Use the table of the standard normal cdf to determine the following probabilities:

$$\begin{aligned} \mathbb{P}(X \leqslant 1) &= \\ \mathbb{P}(X > 1.96) &= \\ \mathbb{P}(X \leqslant -2) &= \\ \mathbb{P}(-1.65 \leqslant X \leqslant 1.65) &= \end{aligned}$$

As the pdf of the standard normal distribution is symmetric around 0,

$$\mathbb{E}\, X =$$

The variance of the standard normal is not so easily calculated, but can be done using integation by parts.

$$\begin{aligned} \mathrm{Var}[X] = \mathbb{E}\,[X^2] - (\mathbb{E}\, X)^2 &= \\ &= && [\text{Take } u = x \text{ and } v = -e^{-x^2/2}] \\ &= \end{aligned}$$

As stated earlier, the normal distribution is a good model for many observed quantities. However, if the normal distribution were restricted to only having expected value zero and variance one, then this would not be the case. Suppose we construct the random variable $Y = \mu + \sigma X$, where $X \sim \mathsf{Normal}(0, 1)$. Then using properties of expectation and variance that we saw earlier

$$\mathbb{E}\, Y =$$
$$\mathrm{Var}(Y) =$$

To determine the distribution of Y we need to study transformations of random variables.

Transformations of a single random variable

Many random variables are constructed by transforming one or more other random variables. Some important examples include:

-
-
-

Linear transformations

In the previous section we constructed a random variable Y from a standard normal random variable X by setting $Y = \mu + \sigma X$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. We saw that $\mathbb{E}\, Y = \mu$ and $\mathrm{Var}(X) = \sigma^2$, but what is the cdf and pdf of Y? For any $y \in \mathbb{R}$,

$$F_Y(Y \leqslant y) = F_X(\mu + \sigma X \leqslant y) =$$

To get the pdf of Y we need to differentiate $F_Y(y)$ with respect to y.

$$\begin{aligned}
\frac{d}{dy} F_Y(Y \leqslant y) &= \frac{d}{dy} F_X(\mu + \sigma X \leqslant y) \\
&= && \text{[apply the chain rule.]} \\
&= \\
&=
\end{aligned}$$

This leads to the following definition the general normal distribution.

Definition. A continuous random variable X is said to have **normal distribution** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ if its probability density function is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

We will denote this by $X \sim \mathsf{Normal}(\mu, \sigma^2)$. The following figure shows the pdf f_X where $X \sim \mathsf{Normal}(\mu, \sigma^2)$, for different parameters μ and σ^2.

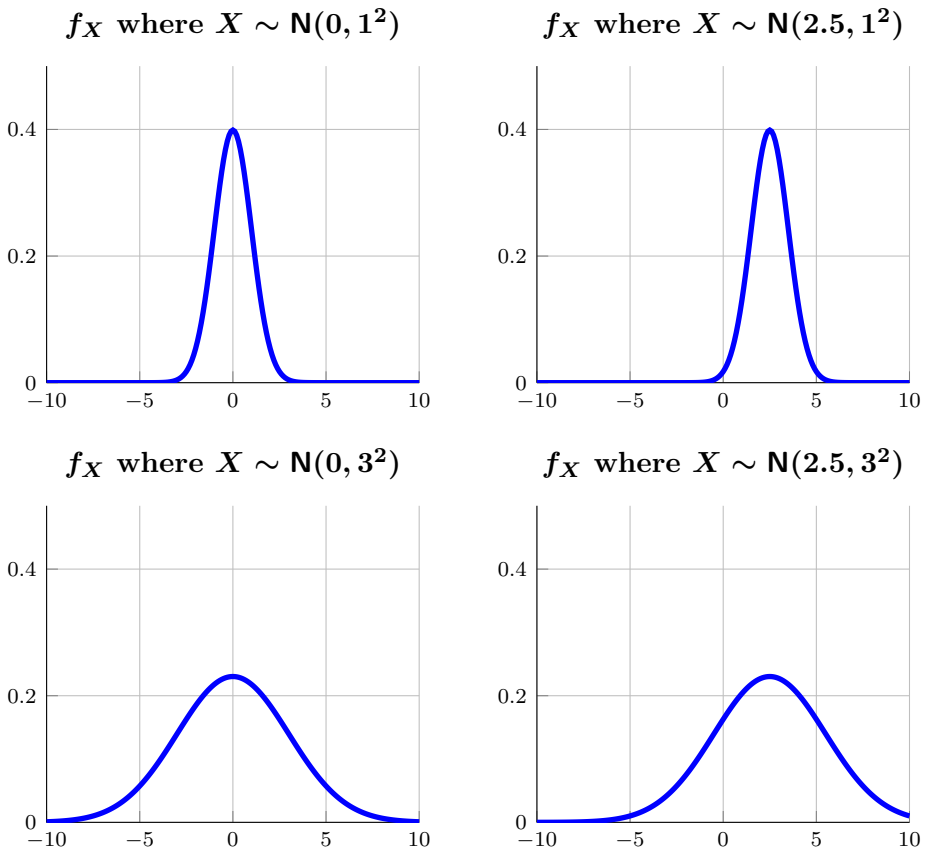


Figure 5.3: The pdfs of $X \sim \mathsf{N}(\mu, \sigma^2)$, with parameters μ and σ^2 given for each plot.

Question. If $X \sim \mathsf{Normal}(\mu, \sigma^2)$, what is the distribution of $Z = (X - \mu)/\sigma$?

$$F_Z(Z \leqslant z) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leqslant z\right)$$
$$=$$
$$=$$
$$\frac{d}{dz}F_Z(z) =$$
$$=$$

So the distribution of Z is $\boxed{}$.

This is an important property that enables us to use the tabulated values of the standard normal cdf to determine the cdf of a normal distribution in general.

Question. Use the tables of the standard normal cdf to determine the following probabilities:

Let $X \sim \mathsf{Normal}(1, 1)$. Find $\mathbb{P}(X \leqslant 1)$.

Let $X \sim \mathsf{Normal}(0, 4)$. Find $\mathbb{P}(X \leqslant 3.92)$.

In general, for a continuous random variable X with pdf $f_X(x)$, the pdf of $Y = aX + b$ is given by

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right),$$

with support $\{y : y = ax + b,\ x \in \mathrm{supp}(X)\}$.

Question. If $X \sim \mathsf{Exponential}(1)$, what is the distribution of $Y = 5X$?

$$f_Y(y) =$$

Monotone transformations

If X is a continuous random variable and $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is monotonic, then we can easily obtain the distribution of Y from that of X. If g is increasing, then, for all $y \in \mathbb{R}$,

$$F_Y(y) = \mathbb{P}(g(X) \leqslant y) =$$

where g^{-1} is the inverse of g. The pdf of Y is then given by

$$\frac{d}{dy}F_Y(y) =$$

If X has support $[a, b]$, then the support of Y is [ ].

On the other hand, if g is decreasing, then

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(X \geqslant g^{-1}(y)) = 1 - \lim_{z \uparrow y} F_X(g^{-1}(z)) = 1 - F_X(g^{-1}(y)),$$

The pdf of Y is then given by

$$\frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - F_X(g^{-1}(y))\right) = \frac{1}{|g'(y)|} f_X(g^{-1}(y)).$$

If X has support $[a, b]$, then the support of Y is [ ].

An important type of monotone transformation is given by the *inverse cdf* or *quantile function*.

Defintion. Let X be a continuous random variable. The function $q_X : \mathbb{R} \to \mathbb{R}$ such that

$$F_X(q_X(x)) = x,$$

is called the **quantile function** of X.

Note that the quantile function is an increasing function.

Suppose F_X is the cdf of a continuous random variable X and q_X is its quantile function. If $U \sim \mathsf{Uniform}(0, 1)$, then the cdf of $q_X(U)$ is

$$\mathbb{P}(q_X(U) \leqslant x) =$$

Example. For $X \sim \mathsf{Exp}(\lambda)$ (consider Figure 5.4) we have:

$$\begin{aligned} F(x) &= 1 - \mathrm{e}^{-\lambda x} \\ &= \\ &= \\ &= \\ \Rightarrow \quad q_X(x) &= -\frac{\ln(1-x)}{\lambda} \end{aligned}$$

when $x \geqslant 0$. Note that if $U \sim \mathsf{Uniform}(0, 1)$, then $V = 1 - U$ has a [ ] distribution.

As a result we can define a MATLAB function to generate samples from the Exponential distribution as follows.

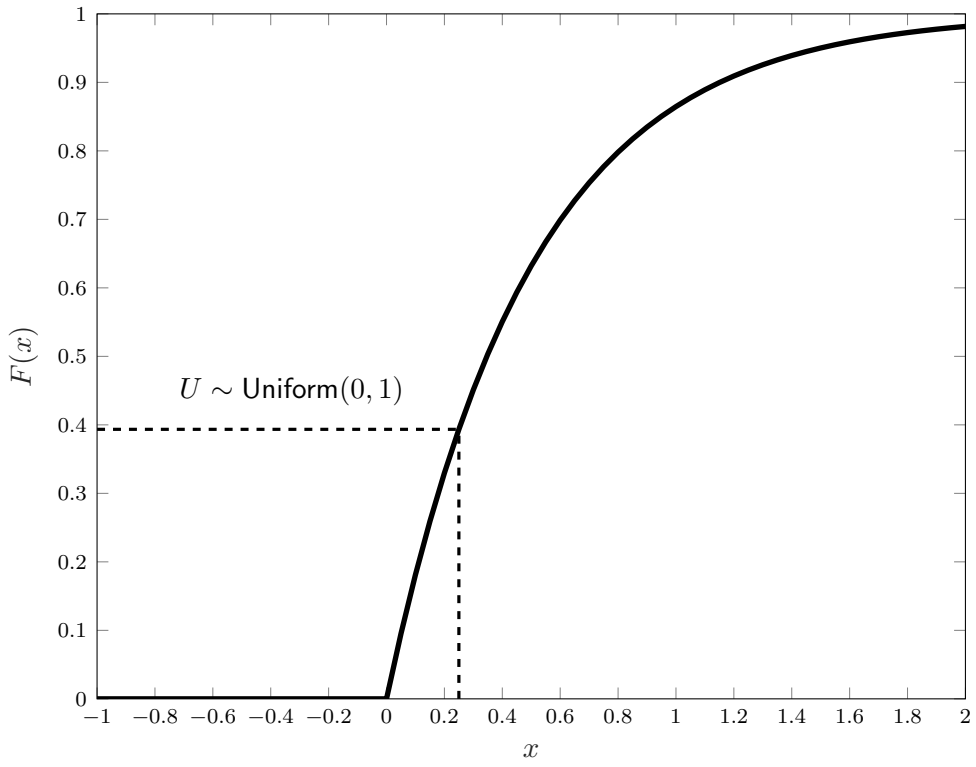


Figure 5.4: The cdf of $X \sim \mathsf{Exp}(2)$.

```
function output = Exponential(lambda)
    output = -log(rand)/lambda;
end
```

Upon saving this as 'Exponential.m' to our working directory we can then use this function as follows:

```
>> Exponential(2)
ans =
    0.0453
>> Exponential(2)
ans =
    0.2291
>> Exponential(2)
ans =
    1.1637
```

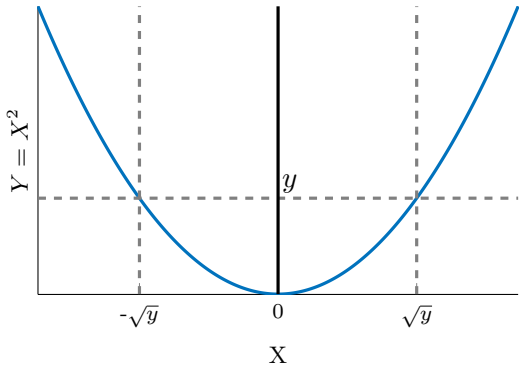
Be careful, as the built in MATLAB function `exprnd` generates samples from the $\mathsf{Exp}(\lambda^{-1})$ distribution.

Non-monotone transformations

If the function g is not monotone, then we can still make progress by considering separately those intervals over which it is monotone. The general procedure to follow is given below:

1. Determine the support of the Y.
2. Determine the event in terms of the random variable X that maps to the event $\{Y \leqslant y\}$. Typically this will be in the form of a union of disjoint events of the form $\{a \leqslant X \leqslant b\}$.
3. Find the probability $F_Y(y)$ of the event $\{Y \leqslant y\}$ in terms of F_X, the cumulative distribution function of X.
4. Differentiate the result to find the probabilty density function of Y.

Example. Suppose $X \sim \mathsf{Normal}(0, 1)$ and $Y = X^2$. Find the pdf of Y.



Step 1: The function $g(x) = x^2$ maps $\mathbb{R}$ to $[0, \infty)$. So the support of Y is $[0\infty)$.

Step 2: From the figure it is clear that $\{Y \leqslant y\}$, where $y \geqslant 0$, corresponds to $\{-\sqrt{y} \leqslant X \leqslant \sqrt{y}\}$.

Steps 3 and 4:

$$\begin{aligned} \mathbb{P}(Y \leqslant y) &= \\ &= \\ &= \end{aligned}$$

so that

$$f_Y(y) =$$

This distribution is called the χ^2_1-distribution.

Example. Suppose $X \sim \mathsf{Exp}(\lambda)$ and $Y = X - \lfloor X \rfloor$. Find the pdf of Y . (*Note that Y is the fractional part of X.*)

Let $y \in (0, 1)$. The event $\{Y \leqslant y\}$ can be written in terms of the random variable X as

$$\mathbb{P}(Y \leqslant y) =$$

$$=$$

$$=$$

so that

$$f_Y(y) =$$

$$=$$

$$=$$

Note. We have previously defined the expected value of a continuous random variable X and the expected value of a function g of X. Now that we have studied the effect of transformations on the distribution of a random variable we can see that these two definitions are consistent. That is, given a continuous random variable X and continouos function g, if we define the random variable $Y := g(X)$, then $\mathbb{E}\, Y = \mathbb{E}\,[g(X)]$.

Multiple continuous random variables

As was the case with discrete random variables, we will often have need to work with multiple random variables at once. Recall that the **joint distribution** of the random variables $X_1, \ldots, X_n$, defined in the same random experiment, can be specified through the **joint cumulative distribution function** F defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) \, .$$

This completely specifies the probability distribution of the vector $\boldsymbol{X} := (X_1, \ldots, X_n)$.

We will now just work with a pair of continuous random variables (X, Y) having joint cdf $F_{X,Y}$. The extension to more than two random variables is straightforward.

Lets first recall some basic notions that we saw previously in connection with the joint distribution of multiple discrete random variables.

It is clear from the law of total probability that

$$F_X(x) = \mathbb{P}(X \leqslant x) = \lim_{y\to\infty} F_{X,Y}(x, y)$$

and similarly for F_Y. We refer to F_X and F_Y as **marginal cumulative distribution functions**. From the joint cumulative distribution $F_{X,Y}$ we can determine if X and Y are independent: X and Y are said to be **independent** if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

for all $(x, y) \in \mathbb{R}^2$. Instead of using the cdf to describe the distribution of a single continuous random variable, we usually used its probability density function. Similarly, for multiple continuous random variables we usually use the joint probability density function.

Definition. If there exists a function $f_{X,Y}(x, y)$ such that for all $(x, y) \in \mathbb{R}^2$

$$F_{X,Y}(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v,$$

we call $f_{X,Y}$ the **joint probability density function** of (X, Y).

Note that in the above double integral we integrate the variable u first, treating v as constant, and then integrate the variable v. In this setting, the order in which we perform this integration is not important since it can be shown that

$$\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\,\mathrm{d}v\,\mathrm{d}u.$$

The joint pdf is not prescribed uniquely by this definition, but, if both of the *partial derivatives* of $F_{X,Y}$ exist at the point (x, y), then

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The symbol $\frac{\partial^2}{\partial x\partial y}$ means to differentiate $F_{X,Y}(x, y)$ first with respect to y, treating x as constant and then differentiate with respect to x, treating y as constant. The order in which we perform this differentiation is not imporant since it can be shown that

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = \frac{\partial^2}{\partial y \partial x} F_{X,Y}(x, y).$$

The joint pdf completely specifies the distribution of (X, Y), as does the joint cdf.

Basic properties of $f_{X,Y}$:

- $f_{X,Y}(x, y) \geqslant 0$ for all $(x, y) \in \mathbb{R}^2$;
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y = 1$.
- $\mathbb{P}(a \leqslant X \leqslant b, c \leqslant Y \leqslant d) = \int_a^b \int_c^d f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$, where a or c can be $-\infty$ and b or d can be ∞, and any of the inequalities can be replaced by strict ones.

Example. Let (X, Y) be a pair of random variables having joint pdf

$$f_{X,Y}(x,y) = \begin{cases} c(x+y+xy), & \text{if } (x,y) = [0,1]^2 \\ 0, & \text{else.} \end{cases}$$

What value must c take for $f_{X,Y}$ to be a valid joint pdf?

Recall that the second axiom of probability states that $\mathbb{P}(\Omega) = 1$, meaning that *something* must happen with probability 1. Here that implies

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,\mathrm{d}u\mathrm{d}v =$$

$$=$$

$$=$$

Therefore, $c =$

The marginal distribution functions F_X and F_Y can be expressed in terms of $f_{X,Y}$; for example,

$$F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,\mathrm{d}v\mathrm{d}u$$

and this leads to the **marginal probability density function** $f_X(x)$ given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,\mathrm{d}v.$$

The marginal pdf f_Y can be similarly determined.

If X and Y are two independent random variables, then their joint pdf can be factorised

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

for all $(x,y) \in \mathbb{R}^2$. The converse is also true so if a joint pdf can be factorised in this way, then the random variables are independent.

Example. Consider again the pair of random variables (X, Y) having joint pdf

$$f_{X,Y}(x,y) = \begin{cases} (x+y+xy), & \text{if } (x,y) = [0,1]^2 \\ 0, & \text{else.} \end{cases}$$

What is the marginal pdf X? Are X and Y independent?

For $x \in [0,1]$, $\quad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,v)\,\mathrm{d}v =$

$=$

Similarly, $\boxed{f_Y(y) =}$ for $y \in [0,1]$. To check independence, for $(x,y) \in [0,1]^2$

$f_X(x) f_Y(y) =$

Similar to the case of a single continuous variable, the expected value of $Z = g(X,Y)$ can be evaluated as

$$\mathbb{E}Z = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$$

and in general, the expected value of $Z = g(X_1,\ldots,X_n)$ can be evaluated as

$$\mathbb{E}Z = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n) f_{\mathbf{X}}(x_1,\ldots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

where $f_{\mathbf{X}}$ is the joint pdf of $(X_1,\ldots,X_n)$.

The properties of expectation that we determined for discrete random variables also hold for continuous random variables. In particular, if X and Y are two continuous random variables measured on the same random experiment, then

$\mathbb{E}[aX+bY] =$ [for any constants $a,b \in \mathbb{R}$]

$\mathbb{E}[XY] =$ [if X and Y are independent]

$\mathrm{Var}(aX+b) =$ [for any constants $a,b \in \mathbb{R}$]

$\mathrm{Var}(X+Y) =$

$\mathrm{Cov}(X,Y) =$ [by definition]

$=$

$\mathrm{Cov}(X,Y) =$ [if X and Y are independent]

$\mathrm{Var}(X+Y) =$ [if X and Y are independent]

Example. Consider again the pair of random variables (X, Y) having joint pdf

$$f_{X,Y}(x,y) = \begin{cases} \frac{4}{5}(x+y+xy), & \text{if } (x,y) = [0,1]^2 \\ 0, & \text{else.} \end{cases}$$

Compute the covariance of X and Y.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathbb{E}[XY] =$$

$$=$$

$$=$$

$$=$$

$$=$$

$$=$$

$$\mathbb{E}\,X =$$

$$=$$

$$=$$

$$\mathbb{E}[Y] =$$

$$\mathrm{Cov}(X,Y) =$$

In general, suppose that $X_1, \ldots, X_n$ are random variables measured on the same random experiment. For arbitrary constants $b_0, b_1, \ldots, b_n$, we have

$$\mathbb{E}\left[b_0 + b_1 X_1 + \cdots b_n X_n\right] =$$

Important: Linearity of expectations holds for any collection of random variables measured on the same random experiment.

Suppose that $X_1, \ldots, X_n$ are *independent* random variables measured on the same random experiment. We have

$$\mathbb{E}\left[X_1 X_2 \cdots X_n\right] =$$

and

$$\mathrm{Var}(X_1 + X_2 + \cdots X_n) =$$

Example. Let Z_1 and Z_2 be two independent standard normal random variables and define $X = Z_1 + Z_2$ and $Y = Z_1 - Z_2$. Compute

$$\mathbb{E}\, X =$$

$$\mathbb{E}\, Y =$$

$$\mathrm{Var}(X) =$$

$$\mathrm{Var}(Y) =$$

$$\mathrm{Cov}(X, Y) =$$

$$=$$

$$=$$

The size of the covariance between X and Y is constrain by the respective variances of X and Y. To see this, we note that for any $t \in \mathbb{R}$,

$$0 \leqslant \mathrm{Var}(X + tY) = \mathrm{Var}(X) + 2t\mathrm{Cov}(X, Y) + t^2\mathrm{Var}(Y).$$

The above quadratic in t must be non-negative for all $t \in \mathbb{R}$. This leads to the inequality

$$|\mathrm{Cov}(X, Y)| \leqslant \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

Definition. The **correlation** (or **correlation coefficient**) of X and Y is defined by

$$\varrho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

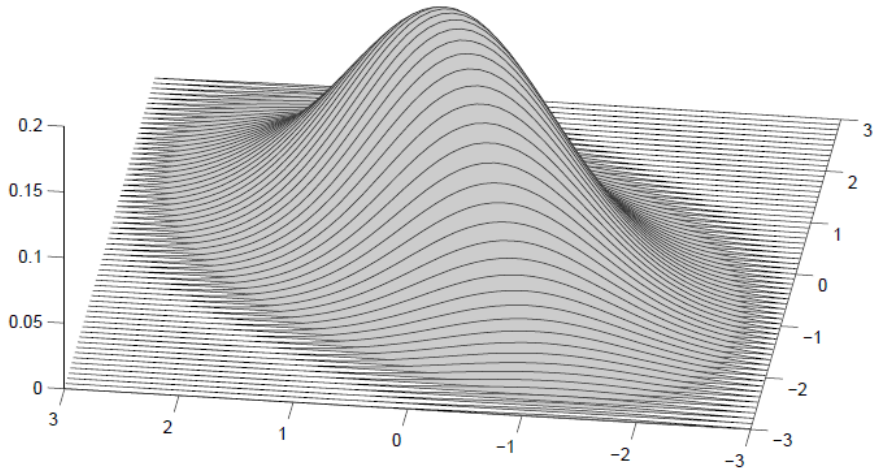
The random variables X and Y are said to be *positively correlated* or *negatively correlated* according as $\varrho(X,Y) > 0$ or $\varrho(X,Y) < 0$; otherwise, they are *uncorrelated*. The larger the value of $|\varrho(X,Y)|$ the more strongly correlated are X and Y.

Multivariate normal (Gaussian) distribution

Let (X,Y) be a pair of random variables with joint probability density function

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left(x^2 - 2\varrho xy + y^2\right)\right), \quad (x,y) \in \mathbb{R}^2,$$

where $\varrho \in (-1,1)$. Below is a plot of this joint pdf with $\varrho = 0.5$.



This is an important model; X and Y are said to have a (standard) *bivariate normal distribution*. To determine the marginal pdfs of X and Y, we first write

$$x^2 - 2\varrho xy + y^2 = (1-\varrho^2)x^2 + (y - \varrho x)^2.$$

Then

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left(x^2 - 2\varrho xy + y^2\right)\right) dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left((1-\varrho^2)x^2 + (y-\varrho x)^2\right)\right) dy \\ &= \\ &= \end{aligned}$$

So marginally X has a $\boxed{}$. Similarly, the marginal distribution of Y is $\boxed{}$. From this we can also see that X and Y are independent (that is, $f_{X,Y}$ factorises as $f_X f_Y$) if and only if $\boxed{}$.

The mean and variance of X and Y are

$$\mathbb{E}\,X = \qquad \mathrm{Var}(X) =$$
$$\mathbb{E}\,Y = \qquad \mathrm{Var}(Y) =$$

We can also evaluate the correlation between X and Y using the same trick that we used to evaluate the marginal pdfs of X and Y.

$$\begin{aligned}
\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y) &= \mathbb{E}\,(XY) \qquad \text{[since the marginals of } X \text{ and } Y \text{ are standard normal]} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left((1-\varrho^2)x^2 + (y-\varrho x)^2\right)\right)\mathrm{d}y\,\mathrm{d}x \\
&= \\
&= \\
&=
\end{aligned}$$

The general bivariate normal distribution is only very slightly more complicated: X and Y are said to have a bivariate normal distribution if its joint pdf $f_{X,Y}(x,y)$ has the form

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\varrho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

The marginal distributions are both normal: $X \sim \mathsf{Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \mathsf{Normal}(\mu_Y, \sigma_Y^2)$. Also ϱ is is the correlation of (X,Y), and X and Y are independent if and only if $\varrho = 0$.

A random vectors $\mathbf{X} := (X_1, \ldots, X_n)$ has a multivariate Normal distribution if the joint pdf has the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where

$$\mathbb{E}(X_i) = \boldsymbol{\mu}_i, \quad \text{and} \quad \text{Cov}(X_i, X_j) = \Sigma_{ij}$$

In particular, if Σ is diagonal, then the $X_1, \ldots, X_n$ are independent random variables with $X_i \sim \mathsf{Normal}(\boldsymbol{\mu}_i, \Sigma_{ii})$.

For us, the most important property of the multivariate Normal distribution is its behaviour under linear transformations.

Suppose $\mathbf{X} := (X_1, \ldots, X_n)'$ has a multivariate Normal distribution. Let $\boldsymbol{a} \in \mathbb{R}^m$ and B is an $(m \times n)$ matrix (with $m \leqslant n$). If $\boldsymbol{X} \sim \mathsf{Normal}(\boldsymbol{\mu}, \Sigma)$, then the random vector $Y := \boldsymbol{a} + B\boldsymbol{X}$ has a $\mathsf{Normal}(\boldsymbol{a} + B\boldsymbol{\mu}, B\Sigma B^T)$.

Example: Suppose that $X_1 \sim \mathsf{Normal}(-1, 2)$ and $X_2 \sim \mathsf{Normal}(1, 3)$ are independent. What is the distribution of $Y = 3 + 2X_1 - X_2$?

Observe

$$Y =$$

and

$$\mathbf{X} \sim$$

so

$$Y \sim$$

Conditional probability density functions and conditional expectation

Recall the definition of conditional probability mass function for discrete random variables;

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x,\, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

provided $f_Y(y) = \mathbb{P}(Y = y) > 0$.

For continuous random variables (X, Y) we can similarly define the **conditional probability density function** of X given $\{Y = y\}$, denoted by $f_{X|Y}(x|y)$, ;

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

when $f_Y(y) > 0$.

Note that if X and Y are independent, then

$$f_{X|Y}(x|y) =$$

when $f_Y(y) > 0$.

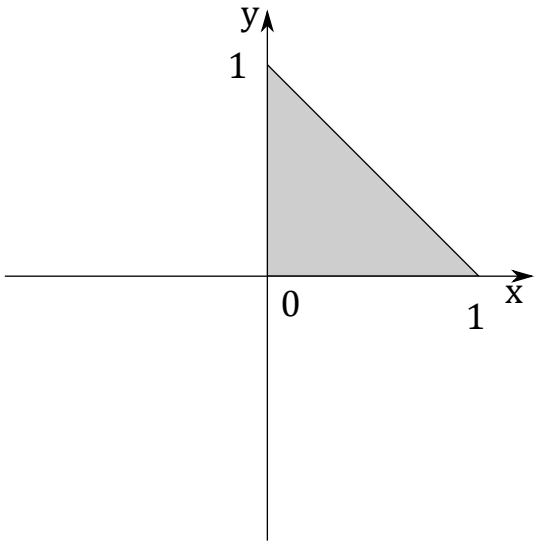
Exercise. Write $f_{X|Y}$ in terms of f_X, f_Y and $f_{Y|X}$.

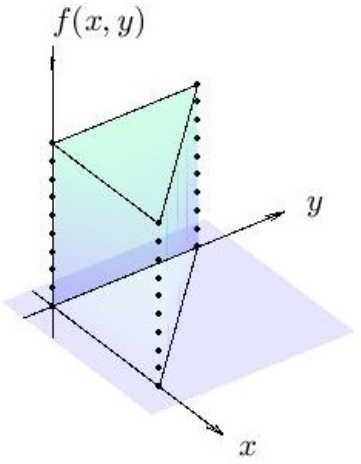
$$f_{X|Y}(x|y) =$$

when $f_Y(y) > 0$.

Example. We draw a random vector (X, Y) uniformly from the triangle $(0,0)$–$(0,1)$–$(1,0)$–$(0,0)$ (see figure). This pdf is only nonzero when both $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1-x$. You can also write those conditions as $0 \leqslant y \leqslant 1$ and $0 \leqslant x \leqslant 1-y$.

What is the joint pdf of X and Y? (Clearly specify where it is zero.)





The triangle has area 1/2. As the joint pdf (of a uniformly-chosen point) must be constant over the support and $\mathbb{P}(\Omega) = 1$, the joint pdf of X and Y is

$$f_{X,Y}(x, y) =$$

What is the marginal pdf of Y and the conditional pdf of X given $\{Y = y\}$ for this example?

The marginal pdf of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x$$

$$=$$

$$=$$

The conditional pdf of X given $\{Y = y\}$ is

$$f_{X|Y}(x|y) =$$

$$=$$

Example. Suppose that (X, Y) has a standard bivariate normal distribution, that is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left(x^2 - 2\varrho xy + y^2\right)\right), \quad (x,y) \in \mathbb{R}^2,$$

where $\varrho \in (-1, 1)$. What is the conditional pdf of Y given $\{X = x\}$?

Using the same trick as before, we can write the joint pdf as

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \frac{1}{\sqrt{2\pi(1-\varrho^2)}} \exp\left(-\frac{1}{2(1-\varrho^2)}(y-\varrho x)^2\right).$$

The marginal distribution of X is $\boxed{}$. So

$$f_{Y|X}(y|x) =$$

$$=$$

That is, conditional on $\{X = x\}$, Y has a $\boxed{}$ distribution.

As we did in the case of discrete random variables, we can take expectations conditional on events such as $\{X = x\}$ or conditional on random variables. We now adapt our definitions to continuous random variables.

Definition. (Conditional expectation given the event $\{X = x\}$) The conditional expectation of Y given $\{X = x\}$ is

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y\, f_{Y\mid X}(y|x)\, \mathrm{d}y.$$

We can relate the conditional expectation $\mathbb{E}[Y \mid X = x]$ to the unconditional expectation $\mathbb{E}[Y]$ as follows: For any random variables X and Y defined in the same random experiement

$$\int_{-\infty}^{\infty} \mathbb{E}\left[Y \mid X = x\right] f_X(x)\,\mathrm{d}x =$$

$$=$$

$$=$$

$$=$$

$$=$$

Example. We saw that if (X, Y) has a standard bivariate normal distribution, then $\mathbb{E}\left[Y \mid X = x\right] = \varrho x$. Using the above property of conditional expectation we see

$$\mathbb{E}\, Y =$$

We do not need to change the definition of conditional expectation given a random variable.

Definition. (Conditional expectation given a random variable) The expression $\mathbb{E}\left[Y \mid X\right]$ is a random variable $g(X)$ that takes the value $\mathbb{E}\left[Y \mid X = x\right]$ when $X = x$.

This conditional expectation has the same properties as in the discrete case.

- Property 1: $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] =$.
- Property 2: Let $Y_1, Y_2, \ldots, Y_n$ and X be random variables defined in the same random experiment. Then $\mathbb{E}\left[\sum_{i=1}^{n} Y_i \middle| X\right] =$
- Property 3: If X and Y are independent, then $\mathbb{E}\left[Y \mid X\right] =$.

Example. Suppose that $X \sim \mathsf{Uniform}(0, 1)$ is chosen, and then $Y \sim \mathsf{Uniform}(X, 1)$. What is the expected value of Y?

The conditional pdf $f_{Y|X}$ is given by

Notice that this conditional pdf is only defined when . For any $< x <$,

$$\mathbb{E}[Y \mid X = x] =$$

$$=$$

$$=$$

$$=$$

$$= \quad .$$

Fortunately, we have the simple formula

$$\mathbb{E}[Y \mid X] = \quad .$$

Now,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]] =$$

$$= \quad .$$

Convolutions

Sums of independent random variables arise regularly in scientific and industrial modelling so it is important to be able to identify its distribution. If X and Y are two independent discrete random variables with support on the integers, then we can determine the probability mass function of $Z = X + Y$:

$$\begin{aligned}\mathbb{P}(Z = n) = \mathbb{P}(X + Y = n) &= \sum_{k=-\infty}^{\infty} \mathbb{P}(X + Y = n | Y = k)\mathbb{P}(Y = k) \\ &= \sum_{k=-\infty}^{\infty} \mathbb{P}(X = n - k | Y = k)\mathbb{P}(Y = k) \\ &= \sum_{k=-\infty}^{\infty} \mathbb{P}(X = n - k)\mathbb{P}(Y = k)\end{aligned}$$

If X and Y are now two independent continuous random variables, there is a similar formula to determine the probability density function of $Z = X + Y$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x, \;\; z \in \mathbb{R}.$$

Hence, the density of Z is the *convolution* of the densities of X and Y.

Example. Suppose X and Y are two independent random variables, each having a $\mathsf{Exp}(\lambda)$ distribution. What is the probability density function of $Z = X + Y$?

For $z \in [0, \infty)$,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x$$

$$=$$

$$=$$

$$=$$

This was not too hard, but what if X had an exponential distribution and Y had a general normal distribution? We could still use the convolution formula, however a simpler approach is to use *moment generating functions*.

Moment generating functions

Let X be a real-valued random variable. The **moment-generating function** of X is the function M_X given by

$$M_X(t) = \mathbb{E}\,\mathrm{e}^{tX},$$

defined for all $t \in \mathbb{R}$ for which the value $\mathbb{E}e^{tX}$ exists (is finite). We will require that $M_X(t)$ exist for all t in some open interval containing the origin.

Example. Let X be a continuous random variable, with pdf f_X. The MGF of X is given by

$$M_X(t) = \qquad\qquad .$$

Example. Suppose that $X \sim \mathsf{Exp}(\lambda)$. The MGF of X is given by

$$M_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x$$

$$=$$

$$=$$

$$= \qquad .$$

Exercise. Show that if $X \sim \mathsf{Normal}(\mu, \sigma^2)$ then the MGF of X is $M_X(t) = \mathrm{e}^{\mu t + \frac{1}{2}t^2\sigma^2}$.

We can first find the MGF of some $Z \sim \mathsf{Normal}(0, 1)$.

$$M_Z(t) =$$

$$=$$

$$=$$

$$=$$

$$=$$

Nifty trick: $X = \mu + \sigma Z$ for some random variable $Z \sim \mathsf{Normal}(0, 1)$, so

$$M_X(t) =$$

$$=$$

$$=$$

$$=$$

$$=$$

The moment generating function of the sum of two independent random variables is determined in a same manner as for discrete random variables.

$$M_{X+Y}(t) = \qquad .$$

Example. Suppose that $X_i \sim \mathsf{Exp}(\lambda_i)$ are independent, for $i = 1, \ldots, n$. The MGF of $\sum_{i=1}^n X_i$ is given by

$$M_{\sum_{i=1}^n X_i}(t) = \qquad .$$

Question. In the above example, what is the MGF $M_{\sum_{i=1}^n X_i}(t)$ when $\lambda_1 = \lambda_2 = \cdots = \lambda_n$?

Exercise. Show that if $X_1 \sim \mathsf{Normal}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{Normal}(\mu_2, \sigma_2^2)$ are independent,

$$aX_1 + bX_2 \sim \mathsf{Normal}\big(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2\big)\,.$$

X_1 and X_2 have MGFs

$$M_{X_1}(t) = \qquad\qquad \text{and} \qquad\qquad ,$$

respectively. The MGF of $Y = aX_1 + bX_2$ is given by

$$\begin{aligned} M_Y(t) &= \\ &= \\ &= \\ &= \end{aligned}$$

showing $Y \sim \mathsf{Normal}\,\big(\mu_Y,\, \sigma_Y^2\big)$.

Whenever $M_X(t)$ is defined, it can be determined from the distribution of X. Is the converse true? Given a moment generating function $M_X(t)$, can the distribution of X be determined? Fortunately, the answer is YES (moment-generating functions have the **uniqueness property**) but proving it requires some complex analysis, so we won't do so in this class.

As a simple example, let's consider an easier task. Suppose that you know $M_X(t) = \mathbb{E}\,\mathrm{e}^{tX}$ explicitly. Can you find the mean $\mathbb{E}\,X$ from this?

.

More generally, given an MGF, we can find the nth moment $\mathbb{E}X^n$ by differentiating $M_X(t)$ n times and then setting $t = 0$. This is where the name *moment-generating function* comes from.

The series expansion $\mathrm{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ (plus the linearity property of expectation) may be used to rewrite the MGF as:

$$M_X(t) = \mathbb{E}\mathrm{e}^{tX} = \mathbb{E}\sum_{n=0}^{\infty} \frac{t^n\, X^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n\, \mathbb{E}X^n}{n!}\,.$$

Note that a necessary (but it turns out, not sufficient) condition for the MGF to exist is that $\mathbb{E}\,X^n < \infty$ for all $n = 1, 2, \ldots$.

Exercise: Write $\mathbb{E}\,X$ and $\mathbb{E}\left[X^2\right]$ in terms of M_X.

$$M_X(t) =$$

$$\Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} M_X(t) =$$

$$\Rightarrow M_X'(0) =$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t) =$$

$$\Rightarrow M_X''(0) =$$

Exercise: Suppose X has moment generating function $M_X(t) = (\lambda/(\lambda - t))^n$, where n is a positive integer. Find the mean and variance of X.

$$M_X(t) =$$

$$M_X'(t) =$$

$$M_X''(t) =$$

$$\mathbb{E}\,(X) =$$

$$\mathrm{Var}(X) =$$

6

Estimation

By the end of this chapter you should:

- Know what a simple random sample is.
- Know the Law of Large Numbers and Central Limit Theorem.
- Be able to compute a confidence interval for a mean and for the difference of two means.
- Be able to compute a confidence interval for a proportion and for the difference of two proportions.

Collections of independent random variables play such an important role in statistics that we give them a special name.

Definition. The random variables $X_1, X_2, \ldots, X_n$ form **simple random sample** of size n if

(a) the X_i's are independent random variables, and

(b) every X_i has the same distribution.

We will often write $\mathbf{X}$ for the simple random sample $X_1, X_2, \ldots, X_n$.

A realisation of a simple random sample forms the *sample data*. We typically denote sample data as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Example: Let X_i be the face value of the ith roll of a fair, six-sided die. Assume that the rolls of the die are independent. The random vector (X_1, X_2, X_3) is a simple random sample of size 3 while an example of the sample data $\mathbf{x} = (x_1, x_2, x_3)$ might be $\mathbf{x} = (5, 1, 5)$.

We have seen a number of families of distributions. Each distribution in the family is determined by a set of parameters. For example, the normal distribution is parameterised by its [ ] and [ ]. We will denote the parameter of a generic distribution by θ. Typically, θ will be unknown and we will need to determine a

suitable single number, based on sample data, that represents an appropriate value for θ.

Definition. A point estimate for θ is a single number $T(\mathbf{x})$ constructed from sample data $\mathbf{x}$ that can be thought of as a sensible value for θ. The random variable $T(\mathbf{X})$, where $\mathbf{X}$ is a random sample, is a point estimator of θ.

It is important to understand how well a given estimator performs. Some basic criteria for judging an estimator are *unbiasedness* and *consistency*.

Definition. We say $T(\mathbf{X})$ is an unbiased estimator of θ if $\mathbb{E}[T(\mathbf{X})] = \theta$.

Example. The estimator $\bar{X}$ corresponding to the sample mean is an unbiased estimator of $\mu := \mathbb{E}[X_i]$ as

$$\mathbb{E}\left[\bar{X}\right] =$$

Example. The estimator S^2 corresponding to the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

is an unbiased estimator of $\sigma^2 := \mathrm{Var}(X_i)$. We saw this in Quiz 7.

Example. Let $X \sim \mathsf{Binomial}(n, p)$. The estimator $\widehat{P} := X/n$ is an unbiased estimator of p.

$$\mathbb{E}\left[\widehat{P}\right] =$$

Definition. The estimator $T(\mathbf{X})$ from a sample of size n is said to be consistent if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|T(\mathbf{X}) - \theta| > \varepsilon\right) = 0.$$

In order to establish consistency of our estimators, it will be useful to learn about the inequalities of Markov and Chebyshev (or is it Tchebycheff, or Tschebyschef, or ...)

Inequalities of Markov and Chebyshev

Consider the function $g_c : [0, \infty) \to [0, \infty)$ defined by

$$g_c(y) := y - c\mathbb{I}(y \geqslant c).$$

Let Y be any non-negative random variables. Then computing the expectation of $g_c(Y)$ shows

$$0 \leqslant \mathbb{E}\left[g_c(Y)\right] = \mathbb{E}\left[Y\right] - c\mathbb{E}\left[\mathbb{I}(Y \geqslant c)\right].$$

So

$$\mathbb{P}\left(Y \geqslant c\right) \leqslant$$

This is called Markov's inequality. Now consider a random variable X with finite mean μ and variance σ^2. Setting $Y = (X - \mu)^2$ and $c = \varepsilon^2$ in Markov's inequality yields

$$= \mathbb{P}\left((X - \mu)^2 \geqslant \varepsilon^2\right) \leqslant$$

This is called Chebyshev's inequality. We can use Chebyshev's inequality to show consistency of estimators. For example, let $\bar{X}$ be the estimator corresponding to the sample mean from a simple random sample of size n. Then

$$\mathbb{P}(|\bar{X} - \mu| \geqslant \varepsilon)) \leqslant$$

From this inequality we can see that $\bar{X}$ is a [ ] estimator of μ. In other words, the average of a large number of independent and identically distributed random variables tends to the expected value as the sample size goes to infinity. This result is known as the **Law of Large Numbers**.

Central limit theorem

If $X_1, \ldots, X_n$ is a simple random sample from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution, then

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim$$

Suppose now we have a simple random sample $X_1, \ldots, X_n$ of size n, where $\mathbb{E}\, X_i = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, but the distribution of the X_i is not necessarily normal. It is one of the remarkable results of probability and statistics that $\bar{X}$ has approximately a normal distribution with mean μ and variance σ^2/n. More precisely, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leqslant x\right) = \Phi(x),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. This result is called the **Central Limit Theorem**.

As a sketch of the ideas involved, suppose that $X_1, \ldots, X_n$ form a simple random sample with $\mathbb{E}\, X_i = 0$ and $\mathrm{Var}(X_i) = 1$. The general result follows by considering random variables $Y_i := \mu + \sigma X_i$. If we can show that the moment generating function $M_{Z_n}(t)$ of $Z_n := n^{-1/2} \sum_{i=1}^n X_i$ converges to $\exp(t^2/2)$ (the MGF of the standard normal distribution) for all t in some neighbourhood of 0, then it follows (*from Lévy's continuity theorem — a result we will not study*) that the distribution of Z_n converges to a standard normal distribution.

Let $M_X(t)$ denote the moment generating function of the X_i. The moment generating function of Z_n is

$$
\begin{aligned}
M_{Z_n}(t) &= \\
&= \\
&= && \text{[by independence]} \\
&=
\end{aligned}
$$

As the moment generating function of the standard normal distribution is $\exp(t^2/2)$, we would like to show that

$$\lim_{n\to\infty} n \ln M_X(tn^{-1/2}) =$$

for all t is some neighbourhood of 0. Let $y = n^{-1/2}$. Then we can write the limit as

$$\lim_{y\to 0} \frac{\ln M_X(yt)}{y^2}.$$

We need to use L'Hopital's rule to evaluate this limit as $\lim_{y\to 0} M_X(yt) =$ for any t in a neighbourhood of 0. Applying L'Hoptial's rule once gives

$$\lim_{y\to 0} \frac{\ln M_X(yt)}{y^2} =$$

This is still of indeterminate form since $M_X'(0) =$. Applying L'Hopital's rule again gives

$$\lim_{y\to 0} \frac{\ln M_X(yt)}{y^2} =$$

as $M_X''(0) =$.

Example. Let $X \sim \mathsf{Binomial}(n, p)$. As $X = X_1 + X_2 + \cdots X_n$, where the X_i are independent $\mathsf{Bernoulli}(p)$ random variables, we have

$$\sim \mathsf{Normal}(0, 1) \quad \text{(approximately)}.$$

Example. Let $X_1, \ldots, X_n$ be a simple random sample from a $\mathsf{Poisson}(\lambda)$ distribution. Define $Y = \sum_{i=1}^n X_i$. We have seen that $Y \sim$. The central limit theorem implies that

$$\sim \mathsf{Normal}(0, 1) \quad \text{(approximately)}.$$

Confidence intervals

How can we gauge the accuracy of an estimator of θ? *Confidence intervals* (sometimes called *interval estimates*) provide a precise way of describing the uncertainty in an estimator.

Ou aim is to construct random variables T_1 and T_2 so that the probability of μ being in the interval (T_1, T_2) is sufficiently high. For example, we might want to construct T_1 and T_2 (with $T_1 < T_2$) that ensure that the probability of the mean μ being in (T_1, T_2) is 95%.

Formally, given random variables $X_1, \ldots, X_n$ whose joint distribution depends on some unknown $\theta \in \Theta$, a **$(1-\alpha)$ stochastic confidence interval** is a pair of statistics

$$T_1(X_1, \ldots, X_n) \quad \text{and} \quad T_2(X_1, \ldots, X_n)$$

with the property that

$$\mathbb{P}(T_1 < \theta < T_2) \geqslant 1 - \alpha\,, \quad \text{for all } \theta \in \Theta\,,$$

where the number $1 - \alpha \in [0, 1]$ is the *coverage probability*.

That is, (T_1, T_2) is a *random* interval, based only on the (as yet to be observed) outcomes $X_1, \ldots, X_n$, that contains the unknown θ with probability at least $1 - \alpha$.

A realisation of the random interval, say (t_1, t_2), is called a **$(1-\alpha)$ numerical confidence interval** for θ.

Remark: Whilst *stochastic* confidence intervals contain the unknown θ with probability at least $1 - \alpha$, their *numerical* counterparts either contain θ or they do not. It may be helpful to think of a Bernoulli analogy, where "success" occurs with probability (at least) $1 - \alpha$ — then outcomes are either "successes" or "failures".

Consider a simple random sample of size n from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. Suppose we know σ^2 and we would like to construct a confidence interval for the unknown parameter μ. We know that

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathsf{Normal}(0, 1).$$

The quantity on the left is often called a *pivot*, because we know its distribution (which does not depend on the unknown parameter of interest) and it contains both a statistic and the unknown parameter of interest.

Hence,

$$\mathbb{P}\left(z_{\alpha/2} \leqslant \frac{\bar{X} -}{\qquad} \leqslant z_{1-\alpha/2}\right) = 1 - \alpha\,,$$

where z_γ is the γ-quantile of the standard normal distribution. For example, a standard normal random variable is contained in the interval $(-1.96, 1.96)$ with probability 0.95.

Rearranging, we have

$$= 1 - \alpha\,.$$

As the standard normal distribution is symmetric about 0, the quantiles satisfy $-z_{\alpha/2} = z_{1-\alpha/2}$.

Hence a stochastic $1-\alpha$ confidence interval for μ in this case is

$$\left(\bar{X} - z_{1-\alpha/2} \quad , \bar{X} + z_{1-\alpha/2} \quad\right),$$

which is often *abbreviated* to

$$\bar{X} \pm z_{1-\alpha/2}$$

So for example, in 95% of the simple random samples from $\mathsf{Normal}(\mu, \sigma^2)$, μ will be within $1.96 \times \sigma/\sqrt{n}$ of $\bar{X}$.

Exercise: Suppose that we wish to determine the average time it takes to write a 2 Gb file to a hard-drive we are testing, as well as to quantify the uncertainty inherent in the estimate. We will assume that write times are Normally distributed, with unknown mean μ but known standard deviation $\sigma = 1$ s. We have the following data:

$$7.2\ s, 8.3\ s, 7.8\ s, 8.1\ s, 7.5\ s\,.$$

Construct a numerical (numerical) 95% confidence interval for the unknown mean.

We calculate

$$\bar{x} =$$

From the tabulated values of the standard normal cdf,

, so

The (numerical) 95% confidence interval is

This is great. We were able to say something about an unknown parameter μ based on our sample. Unfortunately, this is practically useless since there is no reason why we would know what σ^2 is.

Impact of Unknown Variance

For a random sample from a normal distribution with known variance σ^2, we have seen that the estimator corresponding to the *sample mean* $\bar{X}$ is normally distributed. From this we can construct a confidence interval for the unknown mean μ. How can we proceed when σ^2 is unknown?

It is natural to consider replacing σ^2 by the unbiased estimator of σ^2 given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

We will denote the estimate of σ^2 obtained in this way by s^2, that is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

For a simple random sample of size n from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution, we now have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathsf{t}_{n-1},$$

where t_{n-1} denotes the t-distribution with $n-1$ *degrees of freedom*. For each sample size we have a different t-distribution. For small degrees of freedom (that is small sample sizes), the t-distribution has much fatter tails than the standard normal distribution. However, as we increase the degrees of freedom (that is as the sample size increases) the t-distribution converges to the standard normal distribution.

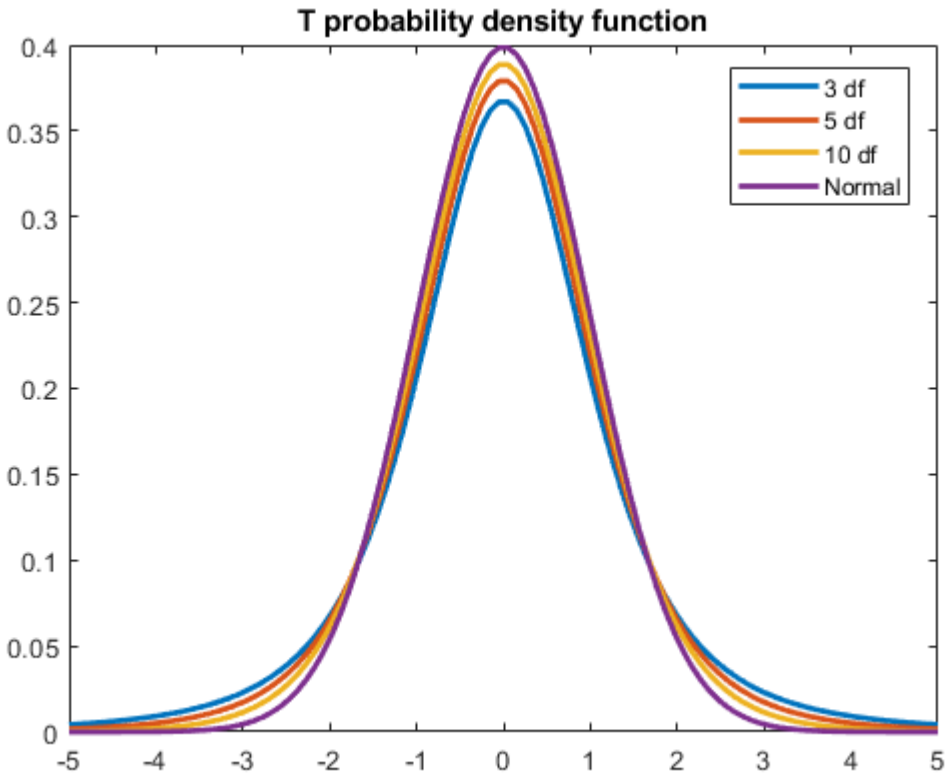


Figure 6.1: The probability density function of the t-distribution with 3, 5 and 10 degrees of freedom together with the standard normal probability density function.

As with the normal distribution, the cumulative distribution function of the t-distribution does not have a simple form and so we will need to refer to tables. There is an added complication since there is a different t-distribution for each sample size. Instead of making a book of tables for the t-distribution only the important (**critical**) values of the t-distribution are tabulated.

With this in mind, lets now construct our confidence interval for μ in the realistic setting

where σ^2 is unknown. We know that

$$\mathbb{P}\left(t_{\alpha/2;\,n-1} \leqslant \frac{\bar{X}-\mu}{S/\sqrt{n}} \leqslant t_{1-\alpha/2;\,n-1}\right) = 1-\alpha\,,$$

where $t_{\gamma;\,n-1}$ is the γ-quantile of the t_{n-1} distribution. Rearranging, we have

$$\mathbb{P}\left(\bar{X} - t_{1-\alpha/2;\,n-1}\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} - t_{\alpha/2;\,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha\,.$$

Like the standard normal distribution, the t-distribution is symmetric about 0 so the quantiles satisfy $-t_{\alpha/2;\,n-1} = t_{1-\alpha/2;\,n-1}$. Hence a stochastic $1-\alpha$ confidence interval for μ when σ^2 is unknown is

$$\left(\bar{X} - t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}}\,,\bar{X} + t_{1-\alpha/2;n-1}\frac{S}{\sqrt{n}}\right).$$

The numerical $(1-\alpha)$ confidence interval, which is what we actually calculate from sample data, is

$$\bar{x} \pm t_{1-\alpha/2;\,n-1}\frac{s}{\sqrt{n}}\,.$$

Some terminology:

- The quantity $s/\sqrt{n}$ is an estimate of the standard deviation of $\bar{X}$. It is called the *standard error* of sample mean and is sometimes denoted by $\mathrm{se}(\bar{x})$.
- The quantity $t_{1-\alpha/2;\,n-1}s/\sqrt{n}$ is the *margin of error* of our estimate $\bar{x}$.

In this section we have worked under the assumption that our simple random sample was from a $\mathsf{Normal}(\mu,\sigma^2)$ distribution. Distributions that arise in practice are rarely exactly normal and so it is important to understand how well our inferential methods perform under deviations from the normal distribution. Due to the central limit theorem, this depends in part on the size of the sample. The general recommendations are as follows:

- For small sample sizes ($n < 15$) we can use methods based on the t-distribution if the data are close to symmetric and there are no *outliers*.
- For moderate sample sizes ($15 \leqslant n < 40$) we can use methods based on the t-distribution as long as there are no outliers or strong skewness in the data.
- For large sample sizes ($40 \leqslant n$) we can use methods based on the t-distribution even in the presence of skewness, though outliers may still affect results.

Let's now consider an example of the material covered in this chapter so far.

Example: Jean-Marc Desharnais[1] surveyed 10 organisations on 81 management information systems development projects completed between 1983 and 1988. Of those projects in the survey, the sample mean for time to completion was 11.7160 months and the sample standard deviation was 7.3997 months. We would like to construct a 95% confidence interval for the mean length of time to completion for a project completed. A histogram of the data is given in Figure 6.2.

[1]The file `deshardnais.xlsx` is on Blackboard. The original data from Desharnais's Masters thesis is available from `http://promise.site.uottawa.ca/SERepository/datasets-page.html`.

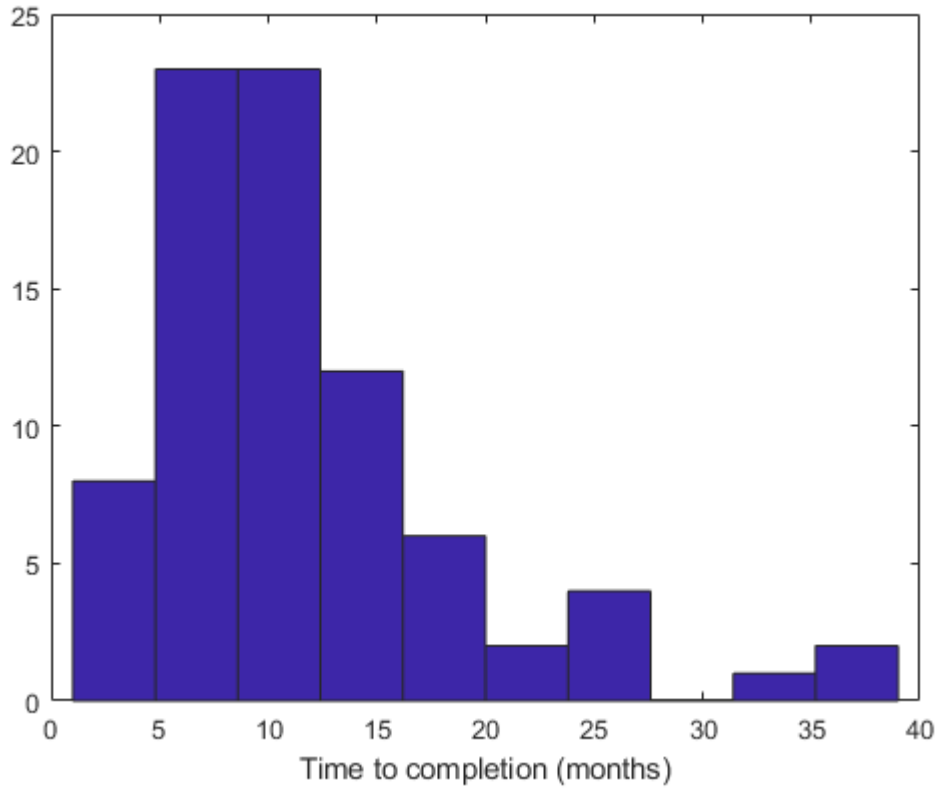


Figure 6.2: Time in months for project completion of 81 management information system development projects.

We are told

$$\bar{x} = \qquad\qquad s = \qquad\qquad n =$$

We want to construct a 95% confindence interval for the mean time to completion μ.

The critical value from the t-distribution in this instance is

$$t_{1-\alpha/2;\,n-1} =$$

The (numerical) 95% confidence interval for μ is

The following code can be used to construct a confidence interval for the mean in MATLAB. First place the file 'desharnais.xlsx' in your current working directory, and then execute the following code.

```
desh = readtable('desharnais.xlsx');
[H,P,CI]=ttest(desh.Length,0,'alpha',0.05);
CI

```

```
5 CI =
6
7     10.0798
8     13.3523
```

The argument `'alpha'` in the function `ttest` corresponds to α giving the $1-\alpha$ coverage probability for the confidence interval.

Difference of two means

Suppose we have two independent simple random samples from two populations. We have m samples $(X_1, \ldots, X_m)$ from the first population which has a $\mathsf{Normal}(\mu_X, \sigma^2)$ distribution and we have n samples $(Y_1, \ldots, Y_n)$ from the second population which has a $\mathsf{Normal}(\mu_Y, \sigma^2)$ distribution. We would like to estimate the difference in the means and be able to quantify our uncertainty about this difference.

To construct a confidence interval for $\mu_X - \mu_Y$, we need to know the distribution of $\bar{X} - \bar{Y}$. We know that $\bar{X}$ and $\bar{Y}$ are independent with $\bar{X} \sim \mathsf{Normal}(\mu_X, \sigma^2/m)$ and $\bar{Y} \sim \mathsf{Normal}(\mu_Y, \sigma^2/n)$. Therefore

$$\bar{X} - \bar{Y} \sim$$

so

$$\frac{(\bar{X} - \bar{Y}) - \qquad\qquad}{\qquad\qquad\qquad} \sim \mathsf{Normal}(0, 1)\,.$$

As before, it is unrealistic to assume that we know σ^2 and so it will need to be estimated. If S_X^2 and S_Y^2 are the sample variance estimators applied to the first and second sample, respectively, then we can form the *pooled variance* estimator by

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

Exercise: Show S_p^2 is an unbiased estimator of σ^2.

$$\mathbb{E}\left[S_p^2\right] =$$

Replacing σ^2 by the pooled variance estimator S_p^2, leads to a statistic that has a t_{m+n-2}-distribution

$$\frac{(\bar{X} - \bar{Y}) - \qquad\qquad}{\qquad\qquad\qquad} \sim t_{m+n-2}\,,$$

Rearranging as usual, we obtain

$$(\bar{X} - \bar{Y}) \pm t_{1-\alpha/2;\, m+n-2}\, S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

or more compactly (and using the symmetry of quantiles of the t-distribution)

$$(\bar{X} - \bar{Y}) \pm t_{1-\alpha/2;\, m+n-2}\, S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

as a $1 - \alpha$ stochastic confidence interval for the difference in means. The numerical confidence interval is

$$(\bar{x} - \bar{y}) \pm t_{1-\alpha/2;\, m+n-2}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}},$$

where

$$s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}.$$

Remark: If our two samples have different variances, then it is still possible to construct an approximate confidence interval for the difference of the two means. The statistic

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$$

has approximately a t-distribution with the degrees of freedom determine by the *Welch* approximation.

Exercise: Suppose that we have two hard-drives of the same model, and we wish to determine an estimate for the average difference in time it takes to write a 2 Gb file to each of the two hard-drives we are testing, as well as to quantify the uncertainty inherent in the estimate. We will assume that write times are Normally distributed, with unknown means μ_X and μ_Y and common variance. We have the following data:

$$\begin{aligned}(x_1, x_2, x_3, x_4, x_5) &= (7.2\ s, 8.3\ s, 7.8\ s, 8.1\ s, 7.5\ s) \\ (y_1, y_2, y_3, y_4, y_5) &= (7.6\ s, 7.3\ s, 8.1\ s, 7.1\ s, 7.0\ s)\,.\end{aligned}$$

Construct a 95% confidence interval for the unknown difference in means, $\mu_X - \mu_Y$.

$\bar{x} =$ $s_x^2 =$ $m =$

$\bar{y} =$ $s_y^2 =$ $n =$

$s_p^2 =$

The critical value we need from the t-distribution is

The (numerical) 95% confidence interval is

This proceedure is implemented in MATLAB with the function `ttest2`.

```
x = [7.2    8.3    7.8    8.1    7.5];
y = [7.6    7.3    8.1    7.1    7.0];
[H,P,CI,STATS] = ttest2(x,y);
CI

CI =

   -0.2873    1.0073

STATS

STATS =

  struct with fields:

    tstat: 1.2824
       df: 8
       sd: 0.4438
```

Confidence intervals for proportions

One of the variables recorded in Desharnais's survey is the number of years experience of the team undertaking the project. Of the 79 projects, 20 were completed by teams with only one year experience. Assuming this survey is representative of the population of information systems development projects, we could construct a confidence interval for the proportion of projects completed by teams with one year experience.

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample with $X_i \sim \mathsf{Bernoulli}(p)$. The natural estimator for p is

$$\widehat{P} = \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

By the central limit theorem,

$$\widehat{P} \sim_{approx} \mathsf{Normal}\,(\qquad,\qquad\qquad)\,,$$

Therefore, we have

$$\mathbb{P}\left(z_{\alpha/2} \leqslant \frac{\widehat{P} -}{} \leqslant z_{1-\alpha/2}\right) \approx 1-\alpha\,.$$

As $\widehat{P}$ is consistent estimator of p, we may replace p in the denominator to obtain

$$\mathbb{P}\left(z_{\alpha/2} \leqslant \frac{\widehat{P} -}{} \leqslant z_{1-\alpha/2}\right) \approx 1-\alpha\,.$$

Rearranging, and using the symmetry of standard normal quantiles, we have

$$\mathbb{P}\left(\widehat{P} - z_{1-\alpha/2} \qquad\qquad \leqslant \quad \leqslant \widehat{P} + z_{1-\alpha/2} \qquad\qquad\right) \approx 1-\alpha\,,$$

which gives as an approximate $1-\alpha$ stochastic confidence interval for p:

$$\widehat{P} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}\,.$$

Returning to our example at the start of this section, suppose we are constructing a 90% confidence interval for proportion of projects completed by teams with one year's experience. Each project will either be completed by a team with one year's experience or by a team with a different amount of experience. Think of the random variables

$$X_i = \begin{cases} 1, & \text{project } i \text{ completed by a team with one year's experience} \\ 0, & \text{else.} \end{cases}$$

Our estimate of this proportion is $\widehat{p} = 20/79 \approx 0.2532$. So the numerical 90% confidence interval for the proportion is

Confidence intervals constructed in this way are only approximate in the sense that the coverage probability for these intervals is appoximately $1-\alpha$. How close the coverage probability is to the desired level depends on how well close the distribution of $\widehat{P}$ is to the normal distribution. A general rule of thumb is that both np and $n(1-p)$ should be at least 10. The MATLAB function `binofit` will produce a confidence interval for a proportion using a different approach to the one described above.

Confidence intervals for a difference of two proportions

Suppose we have two independent simple random samples from two populations. We have m samples $(X_1, \ldots, X_m)$ from the first population which has a $\mathsf{Bernoulli}(p_X)$ distribution and we have n samples $(Y_1, \ldots, Y_n)$ from the second population which has a $\mathsf{Bernoulli}(p_Y)$ distribution. How can we construct a confidence interval for $p_X - p_Y$, the difference of two proportions?

We know that $\widehat{P}_X$ and $\widehat{P}_Y$ are independent with

$$\widehat{P}_X \sim_{approx} \mathsf{Normal}\left(p_X, \frac{p_X(1-p_X)}{m}\right) \quad \text{and} \quad \widehat{P}_Y \sim_{approx} \mathsf{Normal}\left(p_Y, \frac{p_Y(1-p_Y)}{n}\right)$$

so

$$\widehat{P}_X - \widehat{P}_Y \sim_{approx}$$

This leads to the $1-\alpha$ stochastic confidence interval

$$\widehat{P}_X - \widehat{P}_Y \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_X(1-\widehat{P}_X)}{m} + \frac{\widehat{P}_Y(1-\widehat{P}_Y)}{n}}$$

Exercise: A 2014 study aimed to assess the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage. A total of 2,729 women who had positive pregnancy tests at clinics were identified for participation but only 1,061 were ultimately interviewed.

Of the 208 women who were aged 36 years or above (36+), 52 had a miscarriage, while for the remaining 853 women who were under 36 (< 36), 120 had a miscarriage.

Give a 95% confidence interval for the true difference in the rates of miscarriage between women 36+ and women < 36. What does the interval say about the relationship between age and the rate of miscarriage?

7

Hypothesis Testing

By the end of this chapter you should:

- Know how to specify null and alternative hypotheses.
- Be able to apply basic statistical tests.
- Know how to interpret a test statistic and a p-value.
- Be able to understand the types of errors that occur in hypothesis testing.
- Know what factors controls the probability of these errors in hypothesis testing.

In the previous chapter we saw how to estimate basic quantities such as a mean or proportion and how to quantify our uncertainty about those estimates. Another problem that arises in the analysis of data is how to make a decision about our model. This arises naturally in a number of settings:

- Do video games increase aggressive behaviour in children?
-
-
-

Null and Alternative Hypotheses

In statistics, this problem is called a *hypothesis test*. In hypothesis testing, given data, we wish to determine which of two competing hypotheses: the **null hypothesis** (H_0) and the **alternative hypothesis** (H_1).

We begin with a model for the process generating our data. For example, suppose our data is a realisation of a simple random sample (that is, a realisation of a collection of

independent random variables all having the same distribution) from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. In general, we will denote the parameter(s) of the model by θ and the set of all possible parameter values by Θ. We can now specify the null and alternative hypotheses in terms of the parameter θ.

Let Θ_0 and Θ_1 form a partition of the parameter space Θ. That is, $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The null and alternative hypotheses are then specified as

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1.$$

Example: The currently accepted value for the mean density of the Earth is $5.517 g/cm^3$. In 1798 Henry Cavendish presented some observations for the mean density of the Earth. Suppose Cavendish's apparatus produced measurements from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. Potential hypotheses to test would be

- Test $H_0 : \mu = 5.517 g/cm^3$ versus $H_1 : \mu \neq 5.517 g/cm^3$ (two-sided alternative)
- Test $H_0 : \mu = 5.517 g/cm^3$ versus $H_1 : \mu > 5.517 g/cm^3$ (one-sided alternative)
- Test $H_0 : \mu = 5.517 g/cm^3$ versus $H_1 : \mu < 5.517 g/cm^3$ (one-sided alternative)

These two hypotheses are not treated symmetrically. The null hypothesis H_0 is taken as a statement of the "status quo" and we examine the data looking for evidence against H_0.

- If no evidence against H_0 is found, then we accept H_0.
- On the other hand, if evidence against H_0 is found (in the direction of H_1), then we will reject H_0 in favour of the alternative hypothesis H_1.

Test statistics and *p*-values

So before we can decided whether or not to accept the null hypothesis, we need to be able to quantify the evidence against the null hypothesis. We do this using be constructing a **test statistic** and a ***p*-value**.

A test statistic is a function of the data whose distribution under the null hypothesis is known.

Example: Suppose $X_1, \ldots, X_n$ be a simple random sample from $\mathsf{Normal}(\mu, \sigma^2)$ with $\bar{X}$ and S^2 be the usual estimators of μ and σ^2 constructed from the $X_1, \ldots, X_n$. Under the null hypothesis $H_0 : \mu = 5.517 g/cm^3$, the test statistic

$$T(\mathbf{X}) = \frac{\bar{X} - 5.517}{S/\sqrt{n}}$$

has a t_{n-1}-distribution.

When our test statistic computed from the sample data $T(\mathbf{x})$ is 'large' in an appropriate sense, this will indicate evidence against the null hypothesis. This evidence against the null hypothesis is summarised more clearly through the use of a p-value.

- One sided alternative ($H_1 : \theta > \theta_0$) The p-value is given by

  $$\mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})),$$

  where the probability is evaluated under the null hypothesis.

- One sided alternative ($H_1 : \theta < \theta_0$) The p-value is given by

  $$\mathbb{P}(T(\mathbf{X}) < T(\mathbf{x})),$$

  where the probability is evaluated under the null hypothesis.

- Two sided alternative ($H_1 : \theta \neq \theta_0$) The p-value is given by

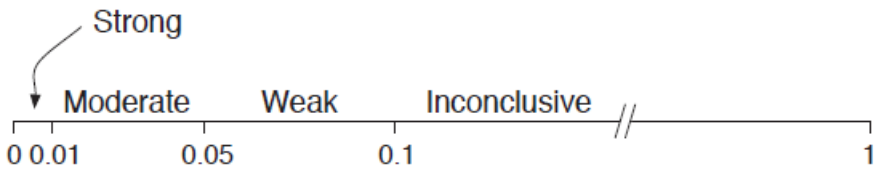
$$2 \min \left[\mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})), \mathbb{P}(T(\mathbf{X}) < T(\mathbf{x})) \right],$$

  where the probability is evaluated under the null hypothesis.

Like the test statistic, the p-value is a function data and so it also has a distribution. Under the null hypothesis

$$p - \text{value} \sim \mathsf{Uniform}(0, 1).$$

The strength of evidence against the null hypothesis provided by the p-value is summarised in the figure below.



We must decide how small the p-value must be before we reject the null hypothesis. This cut-off point is called the **significance level** and is often denoted by α. The significance level determines the probability that we reject the null hypothesis when it is in fact true.

Question: Suppose you were to toss a coin that you believed was fair several times. How many consecutive heads would need to appear before you begin to doubt that it is really a fair coin?

It is common to use significance levels of 5% or 1%, though sometimes much smaller significance levels are needed.

Example: Assume that the measurements from Cavendish's apparatus are a realisation of a simple random sample from $\mathsf{Normal}(\mu, \sigma^2)$. We wish to test whether or not Cavendish's apparatus gave unbiased measurements of the density of the earth, that is we are testing

$$H_0 : \mu = 5.517 g/cm^3 \quad \text{against} \quad H_1 : \mu \neq 5.517 g/cm^3.$$

Cavendish made 23 measurements of the earth's density, with $\bar{x} = 5.4835 g/cm^3$ and $s = 0.1904 g/cm^3$. The test statistic is

$$T(\mathbf{x}) = \frac{\bar{x} - 5.517}{s/\sqrt{n}} = \frac{5.4835 - 5.517}{0.1904/\sqrt{23}} = -0.8438$$

Under H_0, $T(\mathbf{X})$ has a t_{n-1}-distribution. So in this case we need to compare our test statistic with the t_{22}-distribution to get the p-value.

$$
\begin{aligned}
&2\min\left[\mathbb{P}(T_{22} > -0.8438), \mathbb{P}(T_{22} < -0.8438)\right] \\
&= 2\mathbb{P}(T_{22} > 0.8438) \qquad [\text{as } t\text{-distribution is symmetric about zero}] \\
&= 2 \times (0.1,\ 0.25) \qquad [\text{from tables}] \\
&= (0.2,\ 0.5)
\end{aligned}
$$

The p-value is between 0.2 and 0.5. This is inconclusive evidence against H_0. In other words, there is no evidence of bias in Cavendish's apparatus. At the 5% significance level, we retain the null hypothesis.

Connection to confidence intervals

In the previous chapter, we saw how to construct a confidence interval for the mean. Lets now construct a confidence interval for the mean density reading from Cavendish's apparatus.

We can be 95% confident that the mean value of the density measurements made by his apparatus was between 5.401 and 5.566g/cm^3. Note that this interval contains the hypothesised true value of 5.517g/cm^3. Is it just a coincidence that 5.517 was accepted in our hypothesis test?

There is a nice duality between confidence intervals and hypothesis testing. In fact, confidence intervals can be defined as the "inverse" of hypothesis tests:

An alternative definition of a $(1-\alpha)100\%$ confidence interval for a parameter θ is

$$\{\theta \mid \theta \text{ is accepted at } \alpha \text{ significance level (two-sided test)}\}.$$

This is the set of all hypothesised parameter values that would be accepted in a two-sided hypothesis test at significance level α.

Type I and II errors

Whenever we make decisions, we run the risk of making errors. If we reject the null hypothesis when it is in fact true, we have made a **Type I error**. The probability of making a Type I error is precisely the significance level α that we choose for making decisions. For example, if we think a p-value less than $0.05 = 5\%$ is too rare to accept H_0, then we will accidentally reject H_0 precisely 5% of the time.

On the other hand, if we accept H_0 when it is false, then we make a **Type II error**. Related to the notion of Type II errors is the **power** of a statistical test. The power of a statistical test is the probability of detecting an effect when there is indeed an effect. If β is the probability of making a Type II error, then the power is given by $1 - \beta$.

We can think of these errors in terms of a court case:

- A Type I error is accidentally finding someone guilty when they are in fact innocent.
- A Type II error is accidentally finding someone innocent when they are in fact guilty.
- Power is the probability of finding a guilty person guilty.

To summarise, we the following probabilities for all four scenarios:

	Decision	
	Retain H_0	Reject H_0
H_0 is true	Correct $(1-\alpha)$	Type I Error (α)
H_0 is false	Type II Error (β)	Correct $(1-\beta)$

Comparing two means

Example: A real estate agency wants to compare the appraised values of studio apartments in Toowong and Dutton Park. The following results were obtained from random samples:

	Toowong	Dutton Park
Sample Size	25	30
Sample Mean	\$ 226 716	\$ 206 634
Sample Standard Deviation	\$ 32 338	\$ 13 464

Do the two regions have the same (population) mean value for studio apartments?

To address problems like this we follow the same argument that we used to construct the test of a single mean. Suppose we have a simple random sample $X_1, \ldots, X_m$ from a $\mathsf{Normal}(\mu_X, \sigma^2)$ distribution and another simple random sample from $Y_1, \ldots, Y_n$ from a $\mathsf{Normal}(\mu_Y, \sigma^2)$. We want to test the null hypothesis $H_0 : \mu_X - \mu_Y = d$, for some given value d against an alternative hypothesis H_1. The alternative hypothesis is usually one of the following forms:

- One sided alternative: $H_1 : \ \mu_X - \mu_Y > d$.
- One sided alternative: $H_1 : \ \mu_X - \mu_Y < d$.
- Two sided alternative: $H_1 : \ \mu_X - \mu_Y = d$.

Example: For the real estate example, we formulate the null and alternative hypothesis as follows: Let μ_T be the mean appraised value of a studio apartment in Toowong and let μ_D be the mean appraised value of a studio apartment in Dutton Park.

H_0 :	H_1 :

The test statistic for this hypothesis test is

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}},$$

where S_p^2 is the sample pooled variance estimator

$$\begin{aligned} S_p^2 &= \frac{(m-1)S_X^2 + (n-1)S_Y^2}{n+m-2} \\ &= \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{n+m-2}. \end{aligned}$$

As we saw in the previous chapter on confidence intervals, under H_0,

$$T(\mathbf{X}, \mathbf{Y}) \sim t_{n+m-2}.$$

When our test statistic computed from the sample data $T(\mathbf{x}, \mathbf{y})$ is 'large' in an appropriate sense, this will indicate evidence against the null hypothesis. The p-value is given by:

- One sided alternative ($H_1 : \mu_X - \mu_Y > d$) The p-value is given by

$$\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) > T(\mathbf{x}, \mathbf{y})),$$

where the probability is evaluated under the null hypothesis.

- One sided alternative ($H_1 : \mu_X - \mu_Y < d$) The p-value is given by

$$\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) < T(\mathbf{x}, \mathbf{y})),$$

where the probability is evaluated under the null hypothesis.

- Two sided alternative ($H_1 : \mu_X - \mu_Y \neq d$) The p-value is given by

$$2 \min \left[\mathbb{P}(T(\mathbf{X}, \mathbf{Y}) > T(\mathbf{x}, \mathbf{y})), \mathbb{P}(T(\mathbf{X}, \mathbf{Y}) < T(\mathbf{x}, \mathbf{y})) \right],$$

where the probability is evaluated under the null hypothesis.

Example: Lets now perform the test of $H_0 : \mu_T = \mu_D$ against $H_1 : \mu_T \neq \mu_D$. Recall the sample data

	Toowong	Dutton Park
Sample Size	25	30
Sample Mean	\$ 226 716	\$ 206 634
Sample Standard Deviation	\$ 32 338	\$ 13 464

To compute the test statistic we need the pooled variance estimator of σ^2.

$$\begin{aligned} s_p^2 &= \frac{(n_T - 1)s_T^2 + (n_D - 1)s_D^2}{n_T + n_D - 2} \\ &= \frac{24 \times 32338^2 + 29 \times 13464^2}{25 + 30 - 2} \\ &= 5.7274 \times 10^8 \end{aligned}$$

The test statistic is

$$\begin{aligned} T(\mathbf{x}_T, \mathbf{x}_D) &= \frac{(\bar{x}_T - \bar{x}_D) - (\mu_T - \mu_D)}{s_p\sqrt{\frac{1}{n_T} + \frac{1}{n_D}}} \\ &= \frac{(226716 - 206634) - 0}{\sqrt{5.7274 \times 10^8}\sqrt{1/25 + 1/30}} \\ &= 3.0987 \end{aligned}$$

Under the null hypothesis, the test statistic has a t_{53}-distribution. The p-value is

$$\begin{aligned} &2\min\left[\mathbb{P}(T_{53} > 3.0987), \mathbb{P}(T_{53} < 3.0987)\right] \\ &= 2\mathbb{P}(T_{53} > 3.0987) \\ &= 2 \times (0.001,\ 0.005) \qquad \text{[from tables]} \\ &= (0.002,\ 0.01) \end{aligned}$$

This is strong evidence against the null hypothesis in favour of the alternative hypothesis that the mean appraisal value for studio apartments is different for the two regions.

Paired *t*-test

There are situations where we have two samples $(X_1, \ldots, X_n)$ and $Y_1, \ldots, Y_n)$ and although $(X_1, \ldots, X_n)$ are independent and $(Y_1, \ldots, Y_n)$ are independent, X_i and Y_i are dependent for all i. To compare the means of the two populations in this setting, we first take difference $D_i = X_i - Y_i$ and then test the mean of D_i. This often arises when we have two measurements on a single subject before and after some treatment.

Assumptions

As in the previous chapter, we have worked under the assumption that our simple random samples came from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. Our inferences will still hold (approximately) when our samples are not from a normal distribution due to the central limit theorem. The general guidelines are the same as those for confidence intervals.

Test for a single proportion

Suppose you toss a coin 10 times and get 8 heads and 2 tails. Should you be suspicious that this is not a fair coin? Is this evidence that heads are the more likely outcome from the coin toss? We can address this problem using hypothesis testing. Assuming the outcome of each coin toss is independent of the other coin tosses, the number of heads has a $\mathsf{Binomial}(10, p)$ distribution where p is the probability of getting a head on a single coin toss. We want to test the hypothesis:

H_0 : against H_1 :

We want to quantify how surprising it is to get 8 heads from 10 coin tosses which we will do using a p-value. Recall the p-value is the probability of getting data as extreme or more extreme than what we observed assuming the null hypothesis is true. Under the null hypothesis the number of heads has a $\boxed{}$ distribution. So the p-value is

$$p - \text{value} = \mathbb{P}(X \geqslant 8)$$
$$= $$

where $\boxed{X \sim }$ The p-value is 0.0546875. This is moderate evidence against the null hypothesis, suggesting that the coin is biased towards heads.

If we had a two-sided alternative hypothesis, that is $\boxed{H_1 : }$, then the p-value would be computed by

$$p - \text{value} = 2 \min \left[\mathbb{P}(X \geqslant 8), \mathbb{P}(X \leqslant 8) \right].$$

Question: What would we now conclude with the two-sided alternative?

When the number of Bernoulli trials is large, we can use the Central Limit Theorem to evaluate the p-value. Specifically, if $X \sim \mathsf{Binomial}(n, p)$, then

$\dfrac{X - }{} \sim_{\text{approx}} \mathsf{Normal}(0, 1)$

Example. The National Health Interview Survey is conducted annually in the USA by the Center for Disease Control's National Center for Health Statistics. In the 1998-2002 NHIS dataset, there were 25,468 two-child families with children under 10 years old. Of these two-child families 5,844 had two girls. Is there evidence that the proportion of two-child families with two girls is different from 25%?

The null and alternative hypotheses are:

The p-value is

Hence, we conclude ...

The tail probabilities of the binomial distribution can be computed in MATLAB using the function `binocdf`.

Comparing two proportions

In the previous chapter we looked at a 2014 study aimed to assess the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage. There we constructed a 95% confidence interval for the true difference in the rates of miscarriage between women 36+ and women < 36. We now want to formally test if there is any difference in the probability of miscarriage between women 36+ and women < 36. We introduce the following notation:

- let p_1 be the probability that a woman aged 36 years or above has a miscarriage;
- let p_2 be the probability that a woman aged under 36 years has a miscarriage.

We now state our null and alternative hypotheses.

Of the 208 women who were aged 36 years or above (36+), 52 had a miscarriage, while for the remaining 853 women who were under 36 (< 36), 120 had a miscarriage. How can we decide if this represents evidence against the probability of miscarriage in the two groups being equal?

We know that if $X \sim \mathsf{Binomial}(n, p)$, then

$$\widehat{P} = \frac{X}{n} \sim_{approx} \mathsf{Normal}\left(p, \frac{p(1-p)}{n}\right).$$

So if $X_1 \sim \mathsf{Binomial}(n_1, p_1)$ and $X_2 \sim \mathsf{Binomial}(n_2, p_2)$ are two independent random variables, then

$$\widehat{P}_1 - \widehat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2} \sim_{approx}$$

Unfortunately, even under the null hypothesis that $p_1 = p_2$, the (approximate) distribution of $\widehat{P}_1 - \widehat{P}_2$ still depends on the unknown p_1 and p_2.

Let $\widehat{P} = (X_1 + X_2)/(n_1 + n_2)$. Then, assuming the null hypothesis that $p_1 = p_2$ holds,

$$\frac{\widehat{P}_1 - \widehat{P}_2}{\sqrt{\widehat{P}(1-\widehat{P})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{approx} \mathsf{Normal}(0, 1).$$

The (approximate) distribution of this statistic does not depend on unknown parameters so we may use it as our test statistic. Returning to the study on miscarriage, the test statistic is given by

We are testing against the alternative hypothesis $H_1 : p_1 \neq p_2$. The p-value is

Hence, we conclude ...

Contingency tables

The data given in the table below is from the 2014 study on the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage.

	Miscarriage	
Average number of alcoholic drinks per week	Yes	No
4+ drinks per week	11	21
1-3 drinks per week	66	337
No alcohol intake	95	531

Is this data consistent with miscarriage being independent of average number of alcoholic drinks consumed? To answer this question we need to determine what we would expect the data to look like if these two variables were independent. If miscarriage and average alcoholic drinks consumed are independent, then

$\mathbb{P}$ ('Miscarried – Yes' AND 'No alcohol intake) =

We don't know $\mathbb{P}$ ('Miscarried – Yes') or $\mathbb{P}$ ('No alcohol intake), but we could estimate these from the data by

$\widehat{\mathbb{P}}$ ('Miscarried – Yes') = $\widehat{\mathbb{P}}$ ('No alcohol intake) =

So assuming these two variables are independent, the expected number of women in the study who miscarried and had no alcohol intake is

In this way we can create another table containing the expected number of women in each category assuming independence between miscarriage and alcohol consumption.

	Miscarriage	
Average number of alcoholic drinks per week	Yes	No
4+ drinks per week		
1-3 drinks per week		
No alcohol intake		

If the tables of the observed and expected counts differ greatly, then this would be evidence against miscarriage and alcohol consumption being independent. We measure the difference between the two tables using the statistic

$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i}$$

where e_i and o_i are respectively the expected and observed counts for the i-th cell of the table.

In this case, the test statistic evaluates to ...

Assuming that the two variables are independent, X^2 has (approximately) a chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the table. This distribution is often denoted by $\chi^2_{(r-1)(c-1)}$. For this approximation to be reasonable we generally need the expected count for all cells to be at least one and in 80% of cells the expected count is least 5. The χ^2_k distribution has probability density function

$$f(x;k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x > 0,$$

where Γ is the gamma function. As with the normal and t-distributions, the cumulative distribution function of the χ^2_k distribution has been tabulated so we can determine the p-value for the test by looking up the appropriate table:

$$p - \text{value} = \mathbb{P}\left(\chi^2_{(r-1)(c-1)} \geqslant X^2\right).$$

Therefore, the p-value for testing

- H_0: 'Miscarriage' and 'Alcohol consumption' are independent, against
- H_1: There is some association between 'Miscarriage' and 'Alcohol consumption'

is

Hence, we conclude ...

In our example only the number of women in the study was fixed. The proportion of women in each category in the study is reflective of the population from which the sample was taken. An alternative way for a study like this to be conducted is to fix the number of women in each of the groups of alcohol consumption. Denote the probability of miscarriage in the three alcohol consumption groups ('No alcohol', '1-3 drinks per week', '4+ drinks per week') by p_0, p_1 and p_2, respectively. We would then be interested in testing

- H_0: $p_0 = p_1 = p_2$, against
- H_1: at least one of the p_i is different.

Computation of the p-value in this instance proceeds in exactly the same way as for the test of independence. It is only the interpretation of the result which is different. This is sometimes called a test for *homogeneity*.

Exercise: In a study of public attitudes to green roof systems[1], 450 people filled in questionnaires. 66.9% of the respondents replied that they might be interested in installing a green roof on their house. A table recording the respondents age group and interest is given below. Is there a difference in attitudes towards green roof systems across age groups?

Age group	Under 18	18–25	26–40	Over 40	Total
Interested	136	42	48	75	301
Not interested	83	18	31	17	149
Total	219	60	79	92	450

The null and alternative hypotheses are:

[1] Fernandez-Cañero, R., Emilsson, T., Fernandez-Barba, C. & Herrera Machuca, M.A. (2013) Green roof systems: A study of public attitudes and preferences in southern Spain. Journal of Environmental Management, 128, 106-115.

The p-value is ...

Hence, we conclude ...

8

Regression Models

By the end of this chapter you should:

- Be able to perform linear regression or analysis of variance using MATLAB and correctly interpret the output.
- Be able to perform appropriate diagnostic checks.
- Be able to perform appropriate inference for model parameters.

The simplest form of association to analyse between two *quantitative* variables is a linear relationship.

Recall: A variable y is a linear function of x if

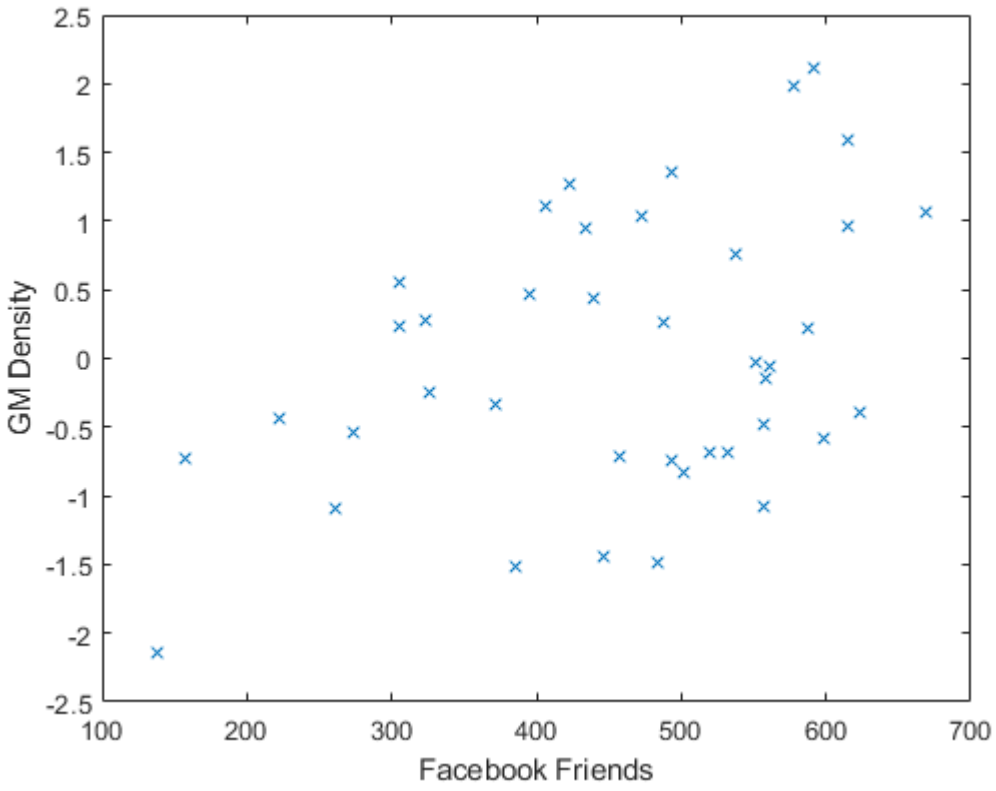
We want to model the relationship between a single random variable Y called the response variable and an explanatory (*as called predictory*) variable x. Our inferences will be based on sample data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, where the $y_1, \ldots, y_n$ are realisations of independent random variables $Y_1, \ldots, Y_n$ however, they do not all have the same distribution. The distribution of Y_i will depend on the explanatory variable x_i.

Connection to two sample inference for means:

Example. Researchers at University College London took a sample of 40 student volunteers and used MRI to measure grey matter (GM) density within three small volumes of the brain. They then looked at the association between these densities and the number of Facebook friends.

The results for one of the volumes, the left middle temporal gyrus, are shown in the following scatterplot. The GM densities are given in standard units.

```
facebook = readtable('facebook.xlsx');
plot(facebook.Facebook,facebook.GMDensity,'x')
xlabel('Facebook Friends')
ylabel('GM Density')
```



In this chapter we model the relationship between two variables as a trend in the mean of the response variable plus variability about that trend.

How would you describe the relationship between grey matter density and the number of Facebook friends?

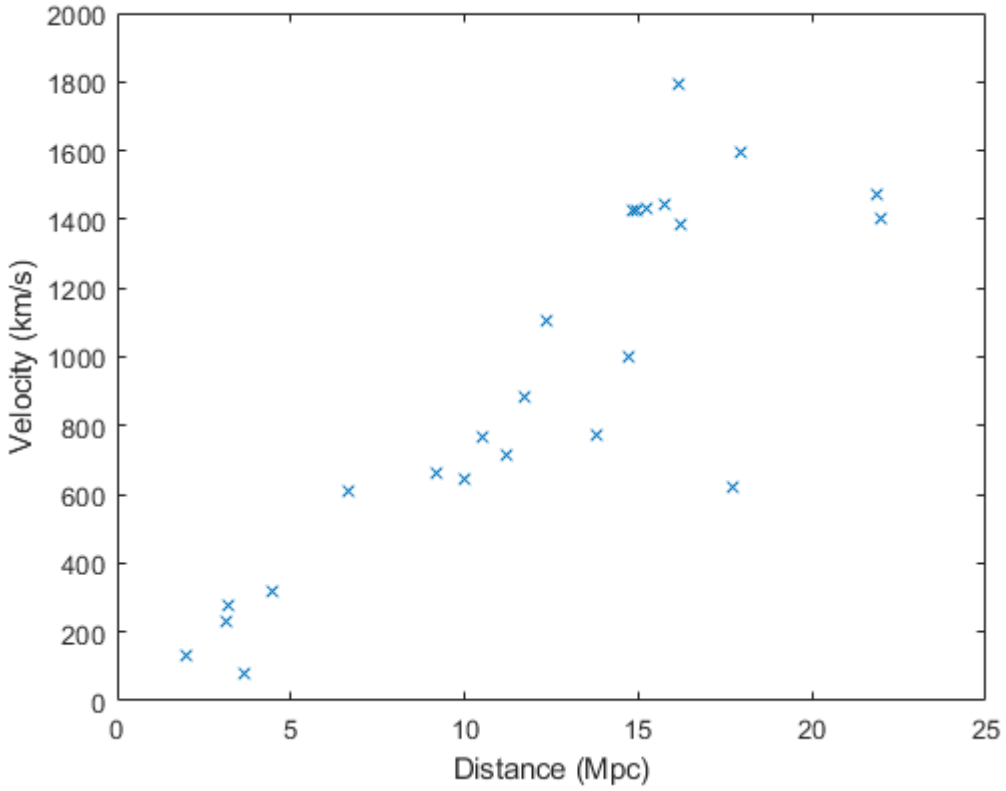
- Direction
- Linearity
- Strength

Example. According to Hubble's law, relative velocity v (km/s) of any two galaxies separated by a distance D (Mega parsec – 1 parsec is 3.09×10^{13} km) is given by

$$v = H_0 D,$$

where H_0 is Hubble's constant. If the expansion of the universe was linear, then $1/H_0$ (Hubble Time) would give the age of the universe. The velocities and distances of 24 galaxies containing Cepheid stars is plotted below (data from the Hubble space telescope key project).

```
hubble = readtable('readtable('hubble.xlsx')');
plot(hubble.x, hubble.y,'x')
axis([ 0 25 0 2000])
xlabel('Distance (Mpc)')
ylabel('Velocity (km/s)')
```



How would you describe the relationship between distance and velocity?

- Direction
- Linearity
- Strength

In the above examples there appears to be a linear trend in the mean of the response variable. In the case of the Hubble data, this is what we expect from the physics. In the case of the Facebook data, there is not theoretical reason why there should be a linear relationship between the number of facebook friends a person has and there grey matter density, but it does appear to describe the data well.

There are usual two main objectives to regression analysis:

- Prediction — Given a new value of the explanatory variable, we would like to predict the value of the response variable with the smallest possible error.
- Explanation — We would like to describe the relationship between the response variable and the explanatory variable.

Supposing we have determined that a straight-line relationship is appropriate, how can we fit a *line of best fit* to the data? In other words, given a line, $b_0 + b_1 x$, how can we judge how well it fits the data?

A widely used measure of fit, and the one we will use in this course, is the sum of squared errors (or sum of squared deviations):

$$\mathcal{S}(\mathbf{b}) = \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2$$

The line that minimises this is called the *least-squares line*.

We can use calculus to find the values of b_0 and b_1 which minimises $\mathcal{S}(\mathbf{b})$. First, we differentiate $\mathcal{S}(\mathbf{b})$ with respect to b_0 and b_1

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_0} =$$

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_1} =$$

Setting these derivatives equal to zero leads to the following system of equations

$$nb_0 + \left(\sum_{i=1}^{n} x_i\right) b_1 = \sum_{i=1}^{n} y_i$$
$$\left(\sum_{i=1}^{n} x_i\right) b_0 + \left(\sum_{i=1}^{n} x_i^2\right) b_1 = \sum_{i=1}^{n} x_i y_i$$

Provided not all the x_i are the same, this system of equations has a unique solution given by

$$b_0 = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$
$$b_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

We (almost) never find the least squares line by hand using these equations. The equations determining the least-squares line are more easily presented in matrix form. Define

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The matrix $\mathbf{X}$ is called the *design matrix*. Then

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \qquad \text{and} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} \\ \\ \end{bmatrix}.$$

We can now write our system of equations in matrix form;

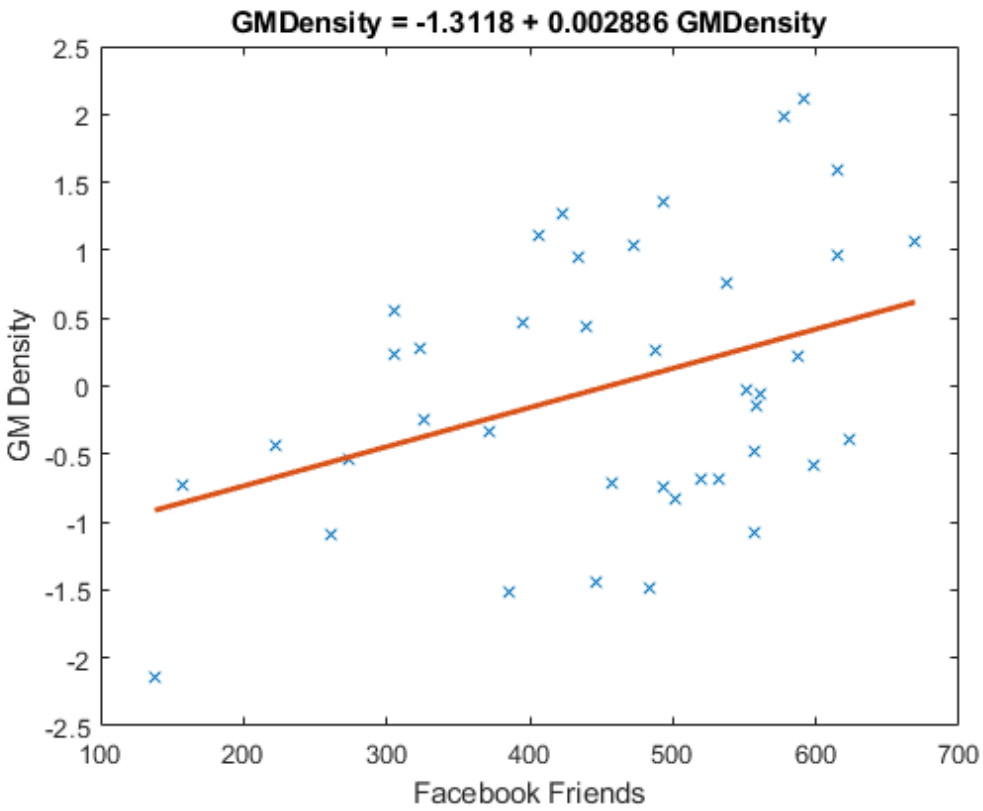
$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$$

whose solution is

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Example. Returning to the facebook example, we can use the `fitlm` function in MATLAB to determine the least-squares line. The least-squares line is plotted with the data in the figure below.

```
fitlm(facebook,'Facebook~GMDensity')
facebooklm.Coefficients.Estimate

ans =

   -1.3118
    0.0029
```



The least-squares line is

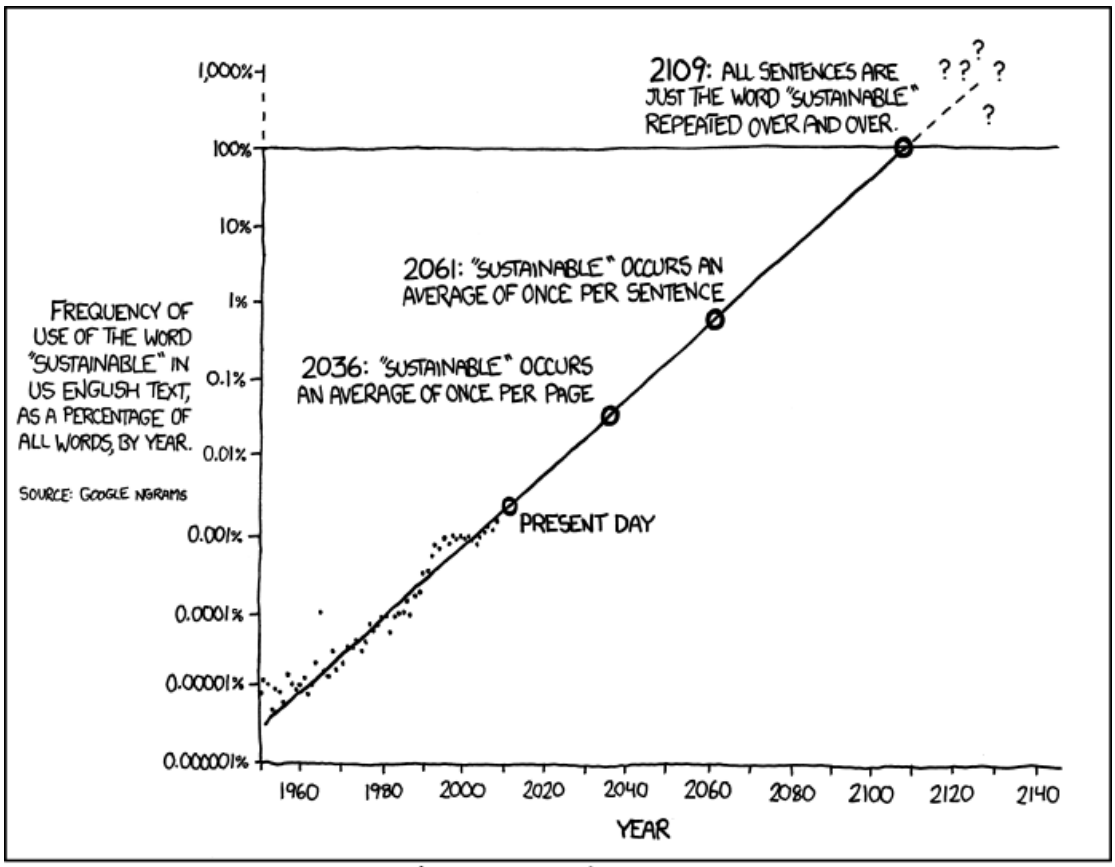
$$-1.311785 + 0.002886 \times \text{Facebook Friends}.$$

How do we interpret the slope and intercept of this line?

- Slope:

- Intercept:

Sensible interpretation of the intercept depends on the context. Extrapolating beyond the range of the data is generally to be avoided.



xkcd.com/1007/

This proceedure allows us to fit a straight-line to our data. However, we know that our data is just the realisation of some random process and so we can think of our estimated line as being the realisation of some random process as well. So, in the facebook example, if we collected data from a different group of 40 students, would we have obtained exactly the same least-squares line?

We need to understand the variation in our fitted straight-line in order to be able to answer questions like:

- How much variability is there in the estimate of the slope?
-
-

To answer these questions we need a model for the process generating our data.

Linear regression model

Our model for the response random variable Y when the explanatory variable is x comprises two components:

- a mean response $\mathbb{E}(Y) = \beta_0 + \beta_1 x$; plus
- variability in the response

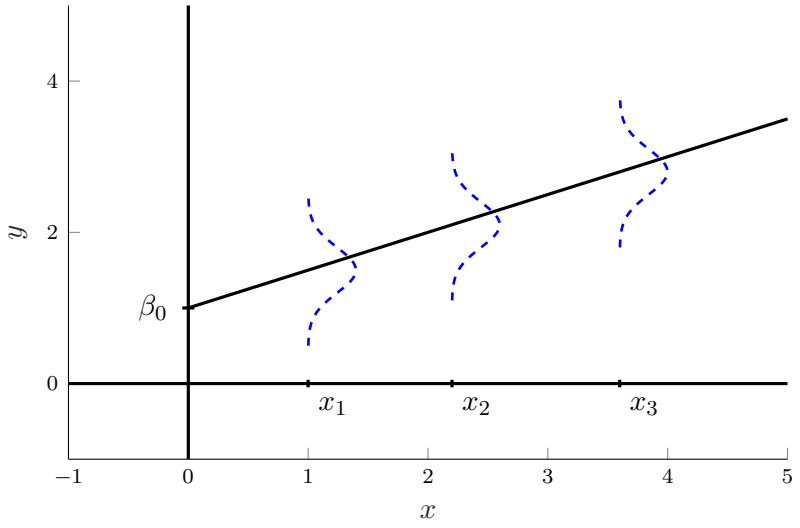
where

- this variability has a Normal distribution
- the amount of variability does not depend on x.

In other words, the response is

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where $\varepsilon \sim \mathsf{Normal}(0, \sigma^2)$, and σ^2 is constant for all x. The (unobservable) errors ε capture deviations from the general trend due to other factors that we did not take into account.



Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where the $Y_1, Y_2, \ldots, Y_n$ are independent random variables and the distribution of Y_i depends on the explanatory variable $x = x_i$. Let $\mathbf{X}$ be the matrix such that $\mathbf{X}_{i,1} = 1$ and $\mathbf{X}_{i,2} = x_i$ and let $\beta = [\beta_0 \quad \beta_1]^T$. We can write the distribution of Y_i and $\mathbf{Y}$ from this model as

$$Y_i \sim \mathsf{Normal}$$

$$\mathbf{Y} \sim \mathsf{Normal}$$

Inference for linear regression

Recall. Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ has a multivariate Normal distribution $\mathsf{Normal}(\boldsymbol{\mu}, \Sigma)$. Let $\boldsymbol{a} \in \mathbb{R}^m$ and B is an $(m \times n)$ matrix (with $m \leqslant n$). Then the random vector $\boldsymbol{a} + B\boldsymbol{Y}$ has a $\mathsf{Normal}(\boldsymbol{a} + B\boldsymbol{\mu}, B\Sigma B^T)$ distribution. In particular, Y_i marginally has a $\mathsf{Normal}(\mu_i, \Sigma_{ii})$ distribution.

Our estimator of the coefficients of the regression line is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

In our linear regression model we assume that $\mathbf{Y} \sim \mathsf{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$. So

$$\mathbf{X}^T\mathbf{Y} \sim \mathsf{Normal}$$

$$\sim \mathsf{Normal}$$

and as $(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \sim \mathsf{Normal}$$

$$\sim \mathsf{Normal}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

Since $\mathbb{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, we say that $\widehat{\boldsymbol{\beta}}$ is an [ ] estimator of $\boldsymbol{\beta}$. Furthermore,

$$\widehat{\beta}_0 \sim \mathsf{Normal}(\beta_0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{11}) \qquad \widehat{\beta}_1 \sim \mathsf{Normal}(\beta_1, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{22})$$

What prevents us from being able to make inferences about $\boldsymbol{\beta}$ at this point is that we don't know σ^2. Our estimator of σ^2 is

$$S^2 = \frac{1}{n-2}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Note that S^2 is an unbiased estimator of σ^2, though we will not try to demonstrate this. So the standard errors of our estimates are

$$s.e.(\widehat{\beta}_0) = s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{11}} \qquad s.e.(\widehat{\beta}_1) = s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{22}}.$$

The main result that we will use to test hypotheses and construct confidence intervals for β is

$$\frac{\widehat{\beta}_i - \beta_i}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{i+1,i+1}}} \sim \mathsf{t}_{n-2}.$$

Example. Returning to the facebook example, when we fitted the linear regression model in MATLAB we got the following output:

```
1 fitlm(facebook,'GMDensity~Facebook')
```

```
facebooklm =


Linear regression model:
    GMDensity ~ 1 + Facebook

Estimated Coefficients:
                    Estimate        SE          tStat      pValue
                    ________    __________    _______    _________

    (Intercept)       -1.3118       0.5397    -2.4306     0.019905
    Facebook        0.0028855    0.0011366     2.5388     0.015341


Number of observations: 40, Error degrees of freedom: 38
Root Mean Squared Error: 0.941
R-squared: 0.145,  Adjusted R-Squared 0.123
F-statistic vs. constant model: 6.45, p-value = 0.0153
```

The `Estimate` column gives the estimate for the intercept term (-1.3118) and the slope of the regression line (0.0028855). The estimate of the slope is reported in the row labelled `Facebook` since this is the name of the explanatory variable in the regression model. The `SE` column reports the standard errors of our estimates of the intercept and slope. Given the estimate of the slope and its standard error we could conduct a hypothesis test

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_1 : \beta_1 \neq 0.$$

For this test, the test statistic has the usual form of

$$\frac{\text{estimate} - \text{hypothesised value}}{s.e.(\text{estimate})}.$$

So our test statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is

$$t =$$

which is, in fact, the value reported in the [ ] column of the output. To get the p-value for this test we compare the test statistic to the $\mathtt{t}_{n-2}$ distribution, where n is the number of observations. There were 40 students in the study (and this is noted in the output) so the degrees of freedom for the $\mathtt{t}-$distribution is [ ] MATLAB also reports this value as the `Error degrees of freedom`. As our alternative hypothesis is two-sided we compute the p-value as

$$\begin{aligned} p - \text{value} &= 2 \times \min\{\mathbb{P}(T_{38} \geqslant 2.5388), \mathbb{P}(T_{38} \leqslant 2.5388)\} \\ &= 2 \times \mathbb{P}(T_{38} \geqslant 2.5388) \end{aligned}$$

This probability can be evaluated in MATLAB using the `tcdf` function

```
2*tcdf(2.5388,38,'upper')

0.015341
```

So p-value for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is actually reported in the `pValue` column of the output.

Question: Does the data provide evidence that a person with no facebook friends has a grey matter density of zero? State the null and alternative hypotheses, and report the appropriate test statistic and p-value from the output. What do you conclude?

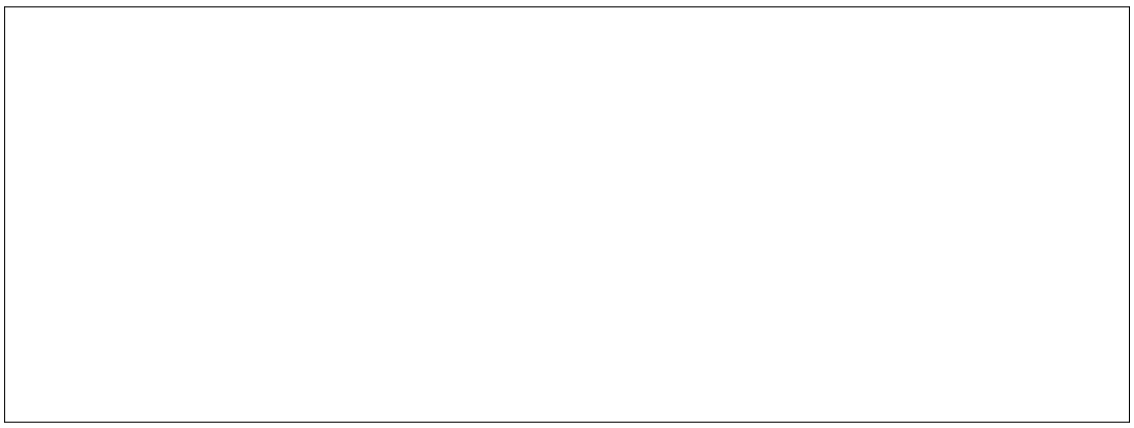
In addition to performing hypothesis tests on the coefficents of the linear regression, we might want to construct confidence intervals for the coefficients. We have the estimates and the corresponding standard errors so a $100(1-\alpha)\%$ confidence interval for true coefficients as

$$\text{estimate} \pm t_{n-2,1-\alpha/2} \times s.e.(\text{estimate}).$$

We can get the critical value from the $\mathtt{t}_{38}$-distribution using MATLAB or tables. For a 95% confidence interval we need $t_{38,0.975}$

```
tinv(0.975,38)

2.024394
```

So the 95% confidence interval for the slope is



MATLAB will compute a desired confidence interval for a coefficient from the linear regression using the function `coefCI`.

Question: Give a 90% confidence interval for the intercept in the linear relationship between grey matter density and the number of facebook friends a person has.

In addition to the coefficients for the regression line β_0 and β_1, the linear regression model also has a parameter σ^2 which is the variance of Y about the mean linear trend. The `fitlm` function of MATLAB returns the estimate σ^2 as

```
Root Mean Squared Error:  0.941.
```

We may like to perform inference for the mean response at a given value of the explanatory variable. For example, we may want to construct a confidence interval for the mean grey matter density for a person who has 250 facebook friends. Let $\mathbf{x}_{new} = [1 \quad x_{new}]$ be the value of the explanatory variable at which we want to construct the confidence interval of the mean response. The true mean response at $\mathbf{x}_{new}$ is $\mathbf{x}_{new}\beta$ and our estimate of this is given by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_{new} = \mathbf{x}_{new}\widehat{\boldsymbol{\beta}}.$$

As we know the distribution of our estimator $\widehat{\boldsymbol{\beta}}$, can find the distribution of estimator of the mean response at the new explanatory variable

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} \sim \mathsf{Normal}(\mathbf{x}_{new}\beta, \sigma^2 \mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}^T).$$

We will use the fact that

$$\frac{\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} - \mathbf{x}_{new}\boldsymbol{\beta}}{S\sqrt{\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}} \sim \mathsf{t}_{n-2}$$

to construct the confidence interval. Using the same kind of reasoning we used in Chapter 6, we arrive at the $100(1-\alpha)\%$ confidence interval

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} \pm t_{n-2;1-\alpha/2} \times s\sqrt{\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}.$$

Returning to the facebook example, lets construct at 95% confidence interval for the mean grey matter density of a person who has 250 facebook friends. So $\mathbf{x}_{new} = [1 \quad 250]$ and

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} = [1 \quad 250] \begin{bmatrix} -1.3118 \\ 0.0028855 \end{bmatrix} = -0.59041.$$

This is our estimate of $\mathbf{x}_{new}\boldsymbol{\beta}$. We now need the standard error of this estimate. We can get the matrix $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ from the result of `fitlm` in MATLAB.

```
1  facebooklm.CoefficientCovariance
2
3   0.2912788549    −5.896487e−04
4  −5.896487e−04     1.291815e−06
```

So

$$\begin{aligned} s^2\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new} &= [1 \quad 250] \begin{bmatrix} 0.2912788549 & -5.896487e-04 \\ -5.896487e-04 & 1.291815e-06 \end{bmatrix} \begin{bmatrix} 1 \\ 250 \end{bmatrix} \\ &= 0.07719289 \\ s.e.(\mathbf{x}_{new}\widehat{\boldsymbol{\beta}}) &= \sqrt{0.07719289} = 0.2778361 \end{aligned}$$

The 95% confidence interval for $\mathbf{x}_{new}\boldsymbol{\beta}$ is

$$-0.59041 \pm t_{38,0.975} \times 0.2778361$$

which is

$$-0.59041 \pm 0.5624498$$

The function `predict` can be used to construct this confidence interval.

```
1  facebooklm = fitlm(facebook,'GMdensity~FacebookFriends');
2  [yhat,ci]=predict(facebooklm,250,'Alpha',0.05)
3
4  yhat =
5
6     −0.5904
7
8
9  ci =
10
11    −1.1529   −0.0280
```

Diagnostics

Just because you can fit a linear regresssion model to your data doesn't mean should. The inferences we make are dependent on the model assumptions. It is necessary to employ some diagnostics to check that our data is consistent with those assumptions.

Recall. The assumptions of the linear regression model are:

- Linearity: The mean of the response variable is a linear function of the explanatory variable.
- Normality: The variability about the mean has a normal distribution.
- Constant variability: The variability about the mean has a constant variance.

The diagnostics we will use are graphical and require some interpretation – and hence experience to use correctly. The diagnostics will be based on the observed *residuals* from the model fit. The residuals are given by

$$\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i.$$

Residuals are not the same as the normal errors ε in the regression model. Note that

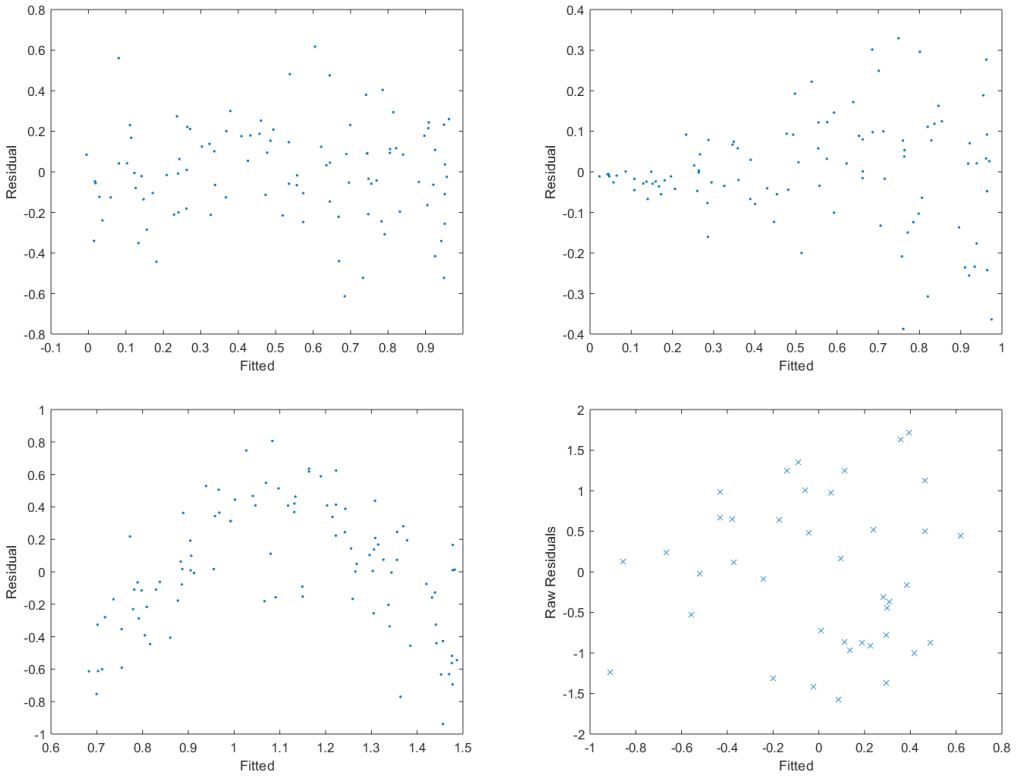
$$\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$=$$

$$=$$

So the residuals will not be independent and not necessarily have constant variance. However, this effect is usally small.

One of the most useful diagnostic plots is the plot of the residuals $\widehat{\varepsilon}$ against the fitted values $\widehat{y} = \mathbf{x}\widehat{\boldsymbol{\beta}}$. In this plot there should be constant symmetrical variation in the vertical direction.



These are plots of the residuals against the fitted values in four regression models. How would you describe these plots in relation to the assumptions of the linear regression model? Are they consistent with the assumptions of Linearity and Constant Variability? (We will use a different plot to check Normality.)

Top-Left:

Top-Right:

Bottom-Left:

Bottom-Right (Facebook data):

To check normality of the residuals we can use the *normal probability plot*. In general, a *probability plot* is used to graphically check that sample data has a specified distribution. The construction of this plot requires the *quantile function*. Recall (from Chapter 5) that the quantile function q for a (continuous) cumulative distribution function F is defined by

$$F(q(x)) = x.$$

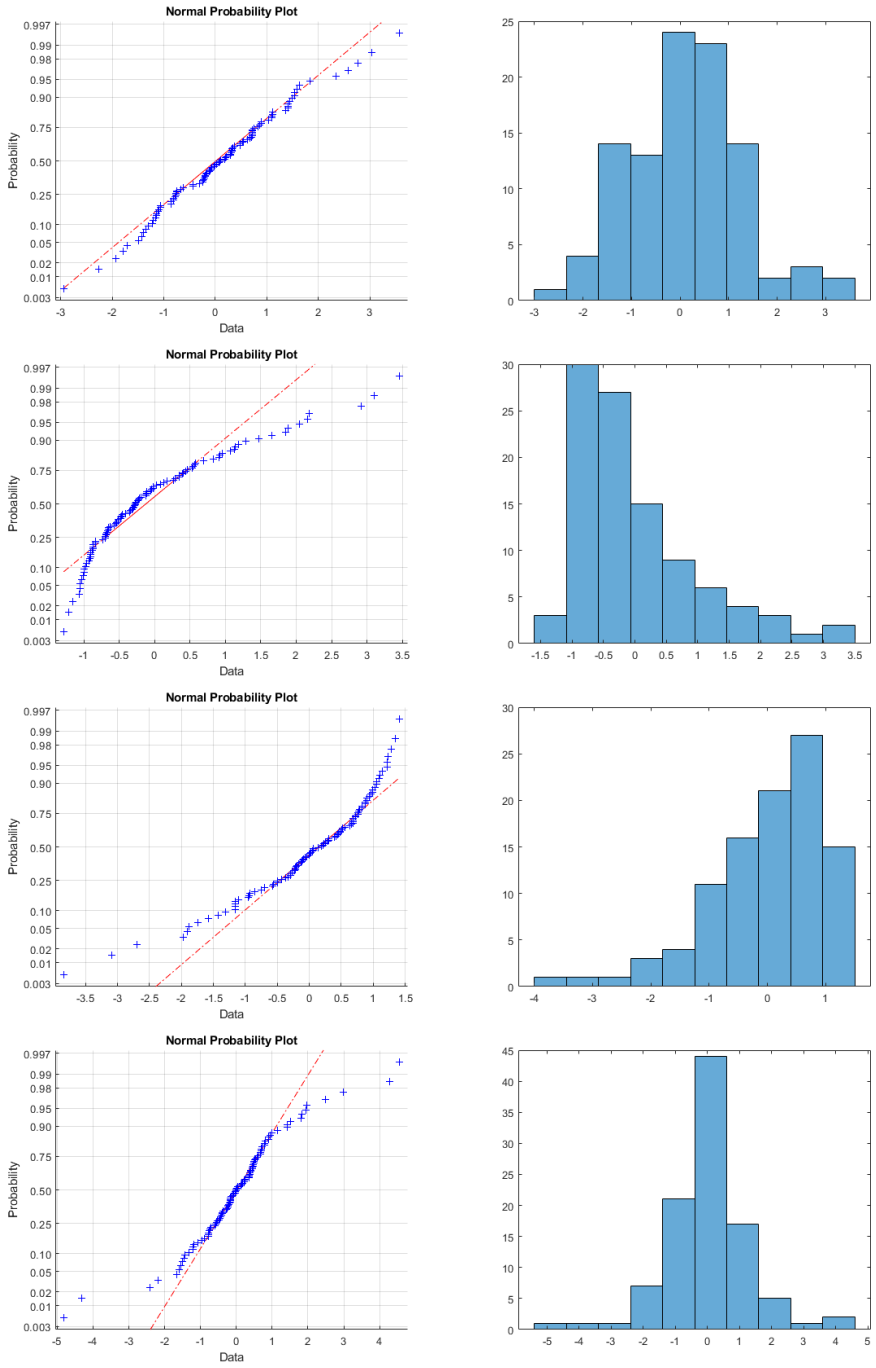
Suppose we have a sample of size n. Then the i-th smallest observation is plotted against the $(i - 0.5)/n$ quantile of the specified distribution. If the observations came from the specified distribution, then the points should lie close to a 45° line. Large departures from a 45° line would suggest that the data are not well modelled by the specified distribution.

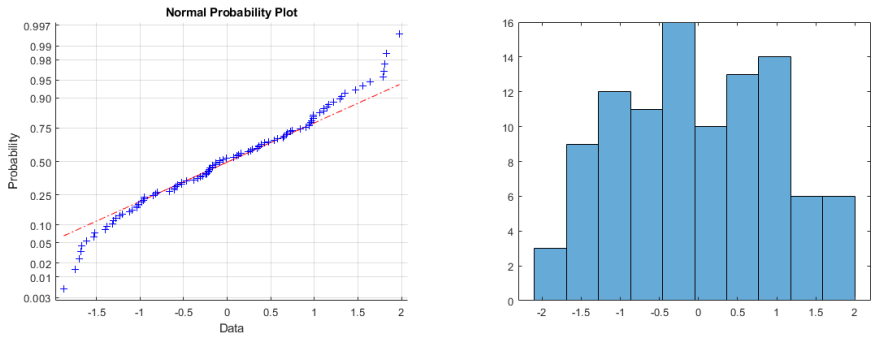
The probability plot, as described, requires the theoretic distribution to be completely specified. This means to check that our observations come from a normal distribution using the probability plot we would need to know the mean and variance of the distribution. However, a nice property of the normal distribution means we can use the quantiles of the standard normal distribution in the probability plot. Let $q(x; \mu, \sigma^2)$ be

the quantile function of the $\mathsf{Normal}(\mu, \sigma^2)$ distribution, then

$$q(x; \mu, \sigma^2) = \mu + \sigma \times q(x; 0, 1).$$

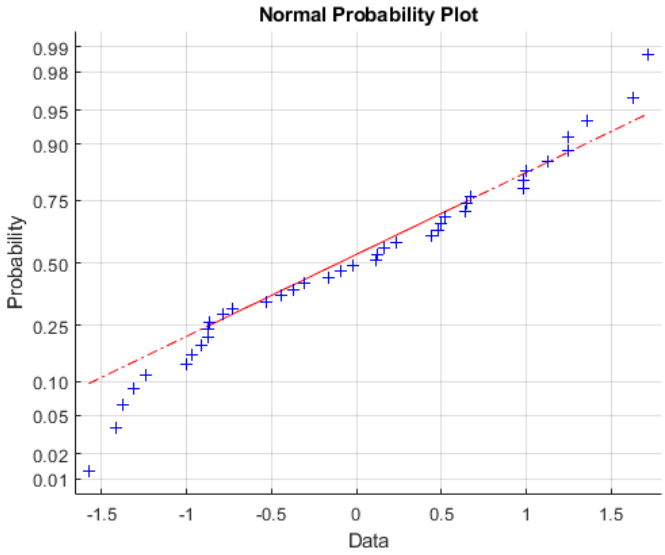
So, if we plot the i-th smallest observation against the $(i-0.5)/n$ quantile of the standard normal distribution and those points lie on a straight line, this would indicate the observations can from a normal distribution. In MATLAB, the y-axis is marked with the probabilities $(i - 0.5)/n$ rather than values of the quantile function, though the spacing from the quantile function is used.





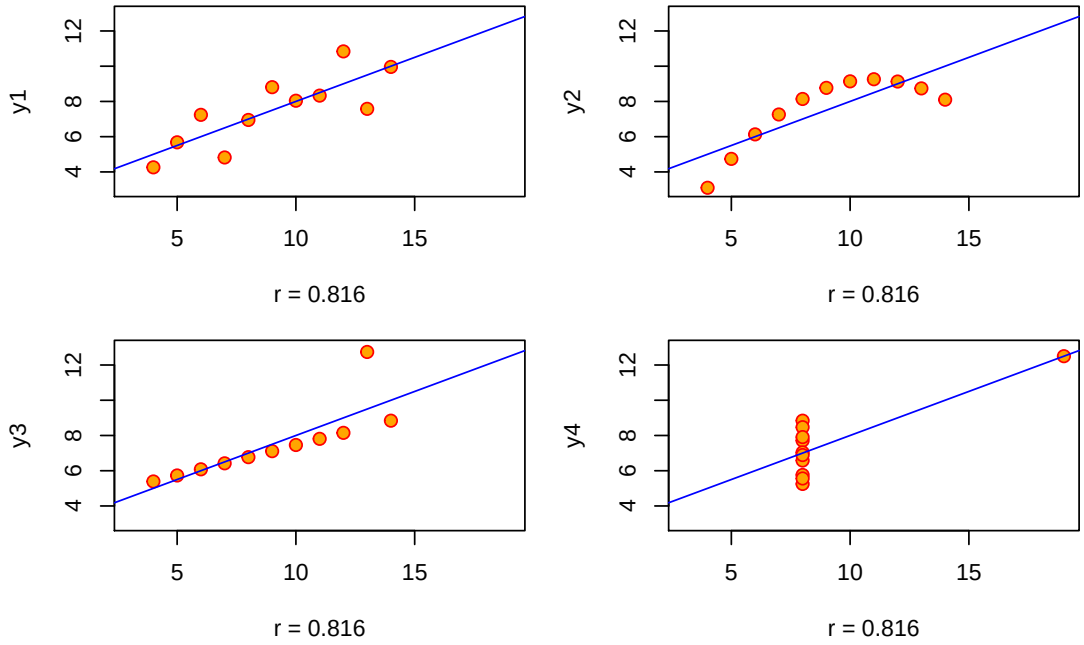
Do the residuals from the facebook example appear normal?

```
facebooklm = fitlm(facebook,'GMDensity~Facebook');
normplot(facebooklm.Residuals.Raw)
```



Independence is another important assumption in our analysis. Unfortunately, there is no simple plot we can do which will tell use whether or not our observations are the realisation of independent random variables. The independence assumption is often called into question when our data has a temporal or spatial aspect.

Question – Anscombe's quartet. Four datasets are plotted below. The exact same linear regression can be fitted to each dataset, but for which, if any, is the linear regression model we have discussed appropriate?

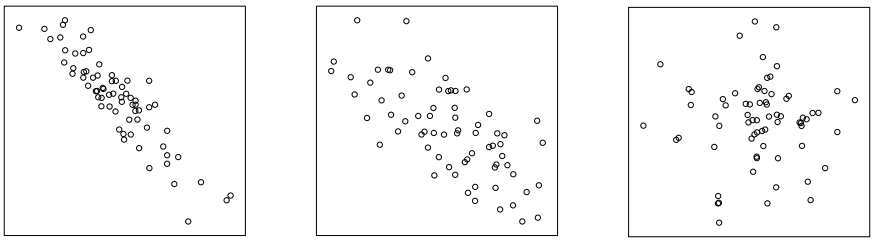


Correlation and the coefficient of determination

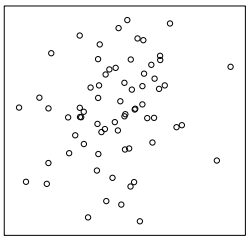
In this chapter we have taken one variable to be the response variable and the other to be the explanatory variable. Sometimes there is no reason to treat the two variables asymmetrically. Instead we might only be interested in the correlation between the two variables. Correlation can be estimated from the pairs of observation using the (Pearson) correlation coefficient, r. If the points in our scatter plot are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, then

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

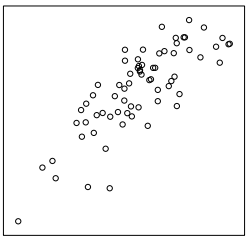
where s_x and s_y are the sample standard deviations of the x and y values of the points. This estimate of correlation is always between -1 and 1.



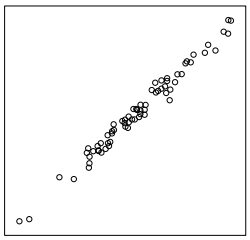
$r = -0.9$ $r = -0.7$ $r = 0$



$r = 0.2$



$r = 0.8$



$r = 0.99$

This is an estimate of the actual correlation between the random variables X and Y:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

We will not derive this, but we can compute the standard error of r as

$$s.e.(r) = \sqrt{\frac{1 - r^2}{n - 2}}.$$

Assuming X and Y are uncorrelated Gaussian random variables, then the statistic

$$\frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} \sim \mathsf{t}_{n-2}.$$

This allows us to test for correlation between X and Y. Let $\varrho = \text{Corr}(X, Y)$. Then we test

H_0: $\varrho = 0$ (there is no correlation between X and Y).

H_1: $\varrho \neq 0$ (there is some correlation between X and Y).

Returning to the facebook example, we might interested in whether there is correlation between the number of facebook friends a person has and their grey matter density. We can use the function `corr` to compute the (Pearson) correlation coefficient. This function will also compute the p-value for testing the null hypothesis $\varrho = 0$ against the alternative hypothesis $\varrho \neq 0$.

```
[rho pval] = corr(facebook.GMDensity,facebook.Facebook)

rho =

    0.3808


pval =

    0.0153
```

Question. Based on this output, what do you conclude? How does your conclusion based on the (Pearson) correlation coefficient compare to your conclusion based on the linear regression model?

Another connection between the (Pearson) correlation coefficient and regression is through the quantity called R^2 (pronounced R-squared). We compute R^2 as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ are the fitted values and $\bar{y}$ is the sample mean of the response variable. We interpret R^2 as the proportion of variation in the response variable that is explained by the regression line. Note that we can re-write our estimates of the regression line coefficients as

$$\widehat{\beta}_1 = r\frac{s_y}{s_x} \qquad\qquad \widehat{\beta}_0 = \bar{y} - \bar{x}r\frac{s_y}{s_x}.$$

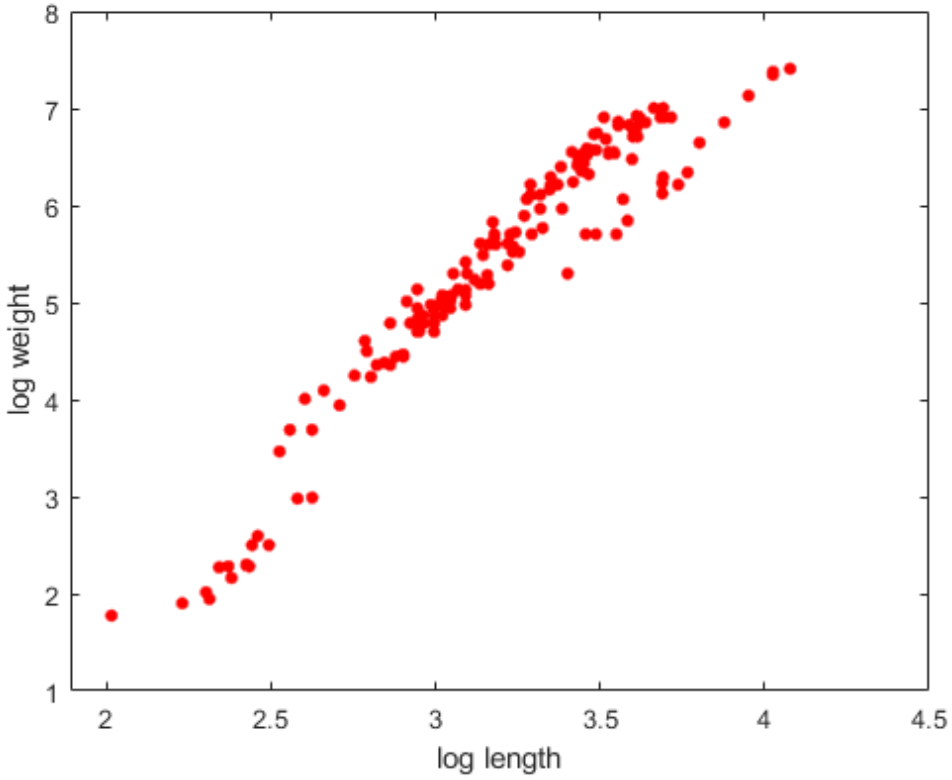
Substituting these expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_0$ into the expression for $\widehat{y}_i$, it can be shown that

$$R^2 = r^2.$$

In other words R^2 is actually the square of the (Pearson) correlation coefficient.

Multiple regression

The weights and lengths of 159 fish caught from the same lake (L´angelm´avesi) near Tampere in Finland. The logarithms of the weights and lengths were taken and plotted below. Geometric considerations lead to us to suspect a linear relationship between log weight and log length.



Question. Does the data look consistent with a linear regression model?

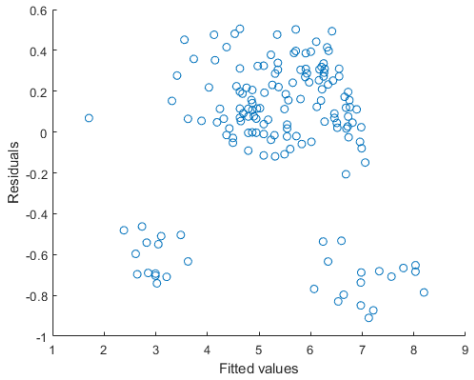
The linear regression model is fitted in MATLAB. The output is reported below.

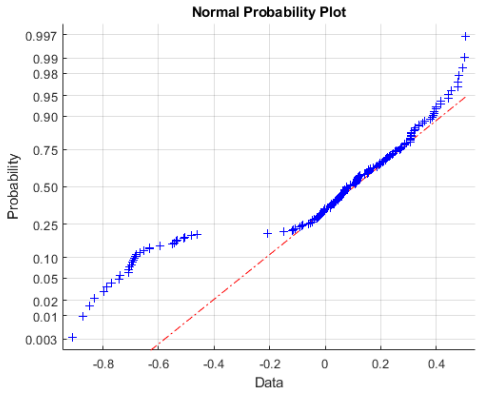
```
Linear regression model:
    logWeight ~ 1 + logLength

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-4.6309	0.23547	-19.667	5.7251e-44
logLength	3.1453	0.073194	42.972	5.214e-88

```
Number of observations: 157, Error degrees of freedom: 155
Root Mean Squared Error: 0.371
R-squared: 0.923,  Adjusted R-Squared 0.922
F-statistic vs. constant model: 1.85e+03, p-value = 5.21e-88
```

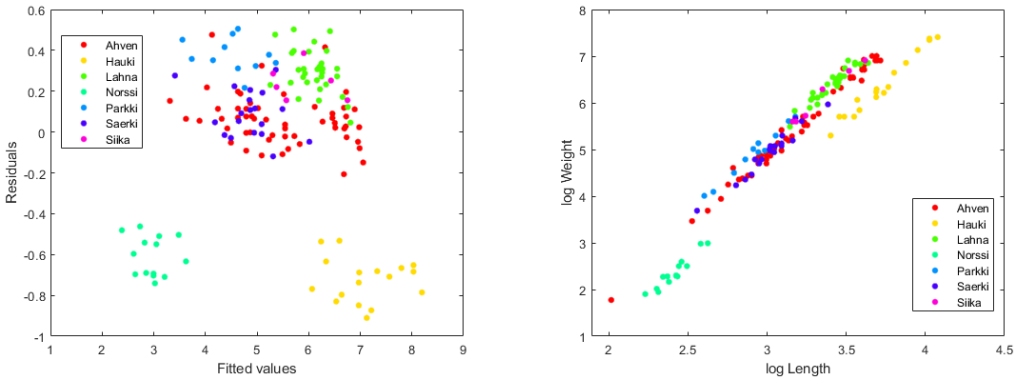
For our inferences (hypothesis tests and confidence intervals) to be valid, the assumption of the model need to hold, at least approximately. We check the model assumptions using diagnostic plots.





Question. Does the data appear consistent with the model assumptions?

The problem is that the fish were not all the same species. There are, in fact, seven different species. In the plot below (left) the residuals are plotted against the fitted values but now the species is indicated. The original log transformed weights and lengths are plotted with the species indicated in the plot on the right.



It seems that the species *Norssi* and *Hauki* require a different intercept term to the other species. Let's concentrate on the *Norssi* species for now. To incorporate this into the linear regression model we need to introduce a *dummy variable* which encodes the species information into a numerical variable.

Let $x_{i,1}$ denote the log length of fish i. We create a new variable $x_{i,2}$ defined as

$$x_{i,2} = \begin{cases} 1, & \text{if the species of fish } i\text{is Norssi} \\ 0, & \text{otherwise} \end{cases}$$

We model the log weight (Y) of a fish as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where $\varepsilon \sim \mathsf{Normal}(0, \sigma^2)$. The mean log weight of a Norssi fish is $\beta_0 + \beta_1 x_1 + \beta_2$ and the mean log weight of other fish species is

We now have three coefficients to estimate. Our least squares estimator of $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2]^T$ still has the form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The resulting estimator still has the multivariate normal distribution

$$\widehat{\boldsymbol{\beta}} \sim \mathsf{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

The estimator of σ^2 changes only slightly to become

$$S^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

where p is the number of columns of $\mathbf{X}$. The main result that we use to test hypotheses and construct confidence intervals for β is

$$\frac{\widehat{\beta}_i - \beta_i}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{i+1,i+1}}} \sim \mathtt{t}_{n-p}.$$

Note the change to the degrees of freedom of the t-distribution. Finally, R^2 is still given by

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where $\widehat{y}_i$ are the fitted values.

Fitting this new model with a separate intercept for the Norssi species gives the following output from MATLAB.

```
Linear regression model:
    logWeight ~ 1 + logLength + Norssi

Estimated Coefficients:
```

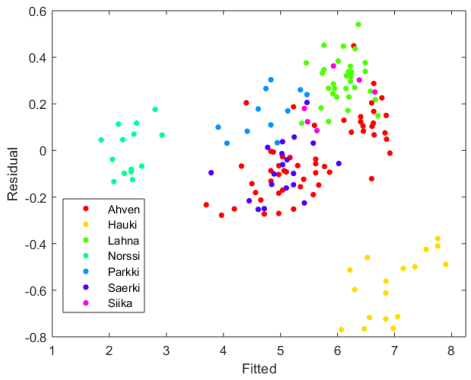
	Estimate	SE	tStat	pValue
(Intercept)	-3.1338	0.23047	-13.598	3.6369e-28
logLength	2.7054	0.070157	38.561	5.0675e-81
Norssi_1	-1.0424	0.099663	-10.459	1.1128e-19

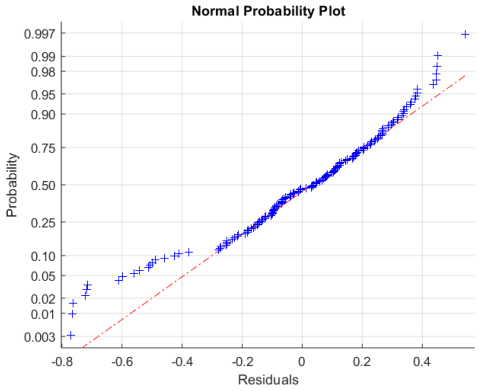
```
Number of observations: 157, Error degrees of freedom: 154
Root Mean Squared Error: 0.285
R-squared: 0.955,  Adjusted R-Squared 0.954
F-statistic vs. constant model: 1.62e+03, p-value = 3.17e-104
```

A couple of points to note relative to the model without the extra intercept term for Norssi fish:

- R^2 has increase from 0.923 to 0.955. As we include more terms in the linear regression model more of the variation in the response variable is explained (R^2 increases). This always occurs.
- In the model with only the log length as the explanatory variable, the p-value reported on the `F-statistic vs constant model` is exactly the same as the p-value reported for the coefficient of `logLength`. In the model with the extra term for Norssi fish, the p-value reported on the `F-statistic vs constant model` is different to the other p-values reported for the coefficients. This p-value relates to the test of hypothesis:
  - $H_0 : \beta_1 = \beta_2 = 0$
  - H_1 : at least one of β_1 and β_2 is non-zero.

Are the assumptions of the linear regression model now satisfied?





We continue until we have a suitable model or decide that the linear regression model is not suitable for our data.

Bibliography

[1] Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability.* CRC Press, Boca Raton, Florida, 2014.

[2] Dirk P. Kroese. A short introduction to probability, 2009.

[3] Dirk P. Kroese and Joshua C.C. Chan. *Statistical Modeling and Computation.* Springer, New York, 2014.

[4] Alberto Leon-Garcia. *Probability, Statistics, and Random Processes for Electrical Engineering.* Prentice Hall, Upper Saddle River, New Jersey, 3 edition, 2007.

[5] David Stirzaker. *Elementary Probability.* Cambridge University Press, Cambridge, 2 edition, 2003.