Contents

1 Probability

1.1 Sample Spaces and Events

1.2 Axioms, Interpretations, and Properties of Probability

1.3 Counting Methods

1.4 Conditional Probability

1.5 Independence

2 Discrete Random Variables and Probability Distributions

2.1 Binomial Distribution

2.1.1 Definition

- Binomial Experiment

  An experiment for which the following 4 conditions are satisfied is called a **binomial experiment**.

  1. The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.
  2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we denote by success (S) or failure (F).
  3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
  4. The probability of success is constant from trial to trial (homogeneous trials); we denote this probability by p.

- Binomial Random Variable

  Given a binomial experiment consisting of n trials, the **binomial random variable** X associated with this experiment is defined as:

$$X = \text{the number of successes among the } n \text{ trials}$$

2.1.2 Notation

We will write $X \sim \text{Bin}(n, p)$ to indicate that X is a binomial rv based on n trials with success probability p. Because the pmf of a binomial rv X depends on the two parameters n and p, we denote the pmf by $b(x; n, p)$. And the cdf will be denoted by $B(x; n, p)$

2.1.3 Formulas

- Probability Mass Function: $b(x;n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$
- Cumulative Distribution Function: $B(x;n,p) = P(X \leq x) = \sum_{y=0}^{x} b(x;n,p) \qquad x = 0, 1, \cdots, n$
- Mean: $E(X) = np$
- Variance: $\text{Var}(X) = np(1-p)$
- Standard deviation: $\sigma_X = \sqrt{npq}$ where $q = 1 - p$

2.2 Poisson Distribution

2.2.1 Definition

A random variable X is said to have a Poisson distribution with parameter λ ($\lambda > 0$) if the pmf of X is

$$p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, \cdots$$

Proposition:
Suppose that in the binomial pmf $b(x; n, p)$ we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\lambda > 0$. Then $b(x; n, p) \to p(x; \lambda)$.

2.2.2 Formulas

- Probability Mass Function: $p(x;\lambda) = \dfrac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, \cdots$
- Cumulative Distribution Function: $B(x; n, p) = P(X \leq x) = \displaystyle\sum_{y=0}^{x} b(x; n, p) \qquad x = 0, 1, \cdots, n$
- Mean: $E(X) = \lambda$
- Variance: $\text{Var}(X) = \lambda$
- Standard deviation: $\sigma_X = \sqrt{\lambda}$.

2.3 The Hypergeometric Distribution

2.3.1 Definition

The assumptions leading to the hypergeometric distribution are as follows:

1. The population or set to be sampled consists of N individuals, objects, or elements (a finite population).
2. Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

The random variable of interest is $X =$ the number of S's in the sample.

2.3.2 Formulas

- Probability Mass Function: $p(X = x) = h(x; n, M, N) = \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
- Mean: $E(X) = n \cdot \frac{M}{N} = np$
- Variance: $\text{Var}(X) = (\dfrac{N-n}{N-1}) \cdot n \cdot \dfrac{M}{N}(1 - \dfrac{M}{N}) = (\dfrac{N-n}{N-1}) \cdot np(1-p)$

2.4 The Neigative Binomial Distribution

2.4.1 Definition

The negative binomial distribution is based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so $P(S \text{ on trial } i) = p$ for $i = 1, 2, 3 \cdots$.
4. The experiment continues (trials are performed) until a total of r successes has been observed, where r is a specified positive integer.

The random variable of interest is $X =$ the number of trials required to achieve the rth success, and X is called a negative binomial random variable.

2.4.2 Formulas

- Probability Mass Function: $nb(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, r+2, \cdots$
- Mean: $E(X) = \frac{r}{p}$
- Variance: $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

2.5 Moment Generating Functions

2.5.1 Definition

- Moment

  The kth moment of a random variable X is $E(X^k)$, while the kth moment about the mean (or kth central moment) of X is $E[(X-\mu)^k]$, where $\mu = E(X)$.
- Moment Generating Functions The moment generating function (mgf) of a discrete random variable X is defined to be

  $$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tx} p(x)$$

  where D is the set of possible X values.

2.5.2 MGF UNIQUENESS THEOREM

If the mgf exists and is the same for two distributions, then the two distributions are identical. That is, the moment generating function uniquely specifies the probability distribution; there is a one-to-one correspondence between distributions and mgfs.

2.5.3 Obtaining Moments from the MGF

If the mgf of X exists, then $E(X^r)$ is finite for all positive integers r, and

$$E(X^r) = M_X^{(r)}(0)$$

PROPOSITION: Let X have the mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{bt} M_X(at)$.

3 Continuous Random Variables and Probability Distributions

3.1 The Normal (Gaussian) Distribution

Definition: A continuous rv X is said to have a normal distribution (or Gaussian distribution) with parameters μ and σ, where $-\infty < \mu < 1$ and $\sigma > 0$, if the pdf of X is

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

The statement that X is normally distributed with parameters μ and σ is often abbreviated $X \sim N(\mu,\sigma)$.

3.1.1 The Standard Normal Distribution

The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution. A random variable that has a standard normal distribution is called a standard normal random variable and will be denoted by Z. The pdf of Z is

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

The cdf of Z is $P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$, which we will denote by $\Phi(z)$.

3.1.2 Non-standardized Normal Distributions

If $X \sim N(\mu,\sigma)$, then the "standardized" rv Z defined by

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

has a standard normal distribution. Thus

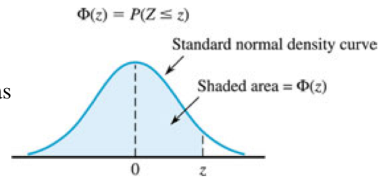
$$P(X \leq a) = \phi(\frac{a-\mu}{\sigma})$$

and the $(100p)$th percentile of the $N(\mu,\sigma)$ distribution is given by

$$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$$

A.3 Standard Normal cdf

Table A.3 Standard normal curve areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3482
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued)

Table A.3 (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

3.1.3 The Normal MGF

The moment generating function of a normally distributed random variable X is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

3.1.4 Approximating the Binomial Distribution

Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. In particular, for $x =$ a possible value of X,

$$\begin{aligned} P(X \leq x) = B(x; n, p) &\approx \text{(area under the normal curve to the left of } x + .5\text{)} \\ &= \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right) \end{aligned}$$

In practice, the approximation is adequate provided that both $np \geq 10$ and $nq \geq 10$.

3.2 The Exponential and Gamma Distribution

3.2.1 The Exponential Distribution

- Definition: X is said to have an exponential distribution with parameter λ ($\lambda > 0$) if the pdf of X is
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$
- Formulas:
  - Cumulative Distribution Function: $F(x;\lambda) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$
  - Mean: $E(X) = \dfrac{1}{\lambda}$
  - Standard deviation: $\sigma_X = \dfrac{1}{\lambda}$

3.2.2 The Gamma Distribution

- Gamma Function: For $\alpha > 0$, the gamma function $\Gamma(\alpha)$ is defined by
$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$
  - Properties:
    1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$ (via integration by parts)
    2. For any positive integer n, $\Gamma(n) = (n-1)!$
    3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
  - Proposition: For any $\alpha, \beta > 0$,
$$\int_0^\infty x^{\alpha-1} e^{-x/\beta}\, dx = \beta^\alpha \Gamma(\alpha)$$
- Gamma Distribution: A continuous random variable X is said to have a gamma distribution if the pdf of X is
$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$
where the parameters α and β satisfy $\alpha > 0$, $\beta > 0$. When $\beta = 1$, X is said to have a **standard gamma distribution**, and its pdf may be denoted $f(x;\alpha)$.
  - The mean and variance of a gamma random variable are
$$E(X) = \mu = \alpha\beta \qquad \mathrm{Var}(X) = \sigma^2 = \alpha\beta^2$$
  - CDF of a standard gamma rv X (Incomplete gamma function)
$$G(x;\alpha) = P(X \leq x) = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy$$
  - Let X have a gamma distribution with parameters α and β. Then for any $x > 0$, the cdf of X is given by
$$P(X \leq x) = G(\frac{x}{\beta};\alpha)$$
the incomplete gamma function evaluated at $\dfrac{x}{\beta}$.

A.4 Incomplete Gamma Function

Table A.4 The incomplete gamma function $G(x;\alpha) = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} dy$

x \ α	1	2	3	4	5	6	7	8	9	10
1	.632	.264	.080	.019	.004	.001	.000	.000	.000	.000
2	.865	.594	.323	.143	.053	.017	.005	.001	.000	.000
3	.950	.801	.577	.353	.185	.084	.034	.012	.004	.001
4	.982	.908	.762	.567	.371	.215	.111	.051	.021	.008
5	.993	.960	.875	.735	.560	.384	.238	.133	.068	.032
6	.998	.983	.938	.849	.715	.554	.394	.256	.153	.084
7	.999	.993	.970	.918	.827	.699	.550	.401	.271	.170
8	1.000	.997	.986	.958	.900	.809	.687	.547	.407	.283
9		.999	.994	.979	.945	.884	.793	.676	.544	.413
10		1.000	.997	.990	.971	.933	.870	.780	.667	.542
11			.999	.995	.985	.962	.921	.857	.768	.659
12			1.000	.998	.992	.980	.954	.911	.845	.758
13				.999	.996	.989	.974	.946	.900	.834
14				1.000	.998	.994	.986	.968	.938	.891
15					.999	.997	.992	.982	.963	.930

3.2.3 The Gamma MGF

The moment generating function of a gamma random variable is

$$M_X(t) = \frac{1}{(1-\beta t)^\alpha} \qquad t < \frac{1}{\beta}$$

3.3 The Beta Distribution

A random variable X is said to have a beta distribution with parameters α, β (both positive), A, and B if the pdf of X is

$$f(x;\alpha,\beta,A,B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (\frac{x-A}{B-A})^{\alpha-1} (\frac{B-x}{B-A})^{\beta-1} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

The case $A = 0$, $B = 1$ gives the **standard beta distribution**.

The mean and variance of X are

$$\mu = A + (B-A) \cdot \frac{\alpha}{\alpha+\beta} \qquad \sigma^2 = \frac{(B-A)^2 \alpha\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

4 Joint Probability Distributions

4.1 Jointly Distributed Random Variable

4.1.1 The Joint Probability Mass Function for Two Discrete Random Variables

Let X and Y be two discrete rvs defined on the sample space $\mathcal{S}$ of an experiment.

- The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

  Note: to prove $p(x, y)$ can be a joint pmf, need to show

  1. $p(x, y) \geq 0$ for all possible pairs of (x, y).
  2. $\sum_X \sum_Y p(x, y) = 1$.

- The **marginal probability mass functions** of X and of Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x, y) \qquad p_Y(y) = \sum_x p(x, y)$$

4.1.2 The Joint Probability Density Function for Two Continuous Random Variables

Let X and Y be continuous rvs.

- $f(x, y)$ is the **joint probability density function** for X and Y if for any two-dimensional set A,

$$P((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy$$

  In particular, if A is the two-dimensional rectangle $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, then

$$P((X, Y) \in A) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\, dy\, dx$$

- The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \qquad \text{for } -\infty < x < \infty$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \qquad \text{for } -\infty < y < \infty$$

4.1.3 Independent Random Variables

Two random variables X and Y are said to be **independent** if for every pair of x and y values,

$$p(x, y) = p_X(x) \cdot p_Y(y) \text{ when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \text{ when } X \text{ and } Y \text{ are continuous}$$

If this equation is not satisfied for all (x, y), then X and Y are said to be **dependent**.

- Independence of two discrete random variables X and Y requires that every entry in the joint probability table be the product of the corresponding row and column marginal probabilities.
- Independence of two continuous random variables X and Y requires that $f(x, y)$ must have the form $g(x) \cdot h(y)$ and the region of positive density must be a rectangle whose sides are parallel to the coordinate axes.
- For two independent random variables X and Y,

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b) \cdot P(c \leq Y \leq d)$$

4.1.4 More Than Two Random Variable

If $X_1, X_2, \cdots, X_n$ are all discrete random variables, the **joint pmf** of the variables is the function

$$p(x_1, x_2, \cdots, x_n) = P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n)$$

If $X_1, X_2, \cdots, X_n$ are all continuous random variables, the **joint pdf** of the variables is the function $f(x_1, x_2, \cdots, x_n)$ such that for any n intervals $[a_1, b_1], \cdots, [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, \cdots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \cdots, x_n)\, dx_n \cdots dx_1$$

4.2 Expected Values, Covariance, and Correlation

4.2.1 Expected Value

A weighted average of $h(X, Y)$ where the weight functions are $p(x, y)$ or $f(x, y)$ of X and Y.

- Definition

  Let X and Y be jointly distributed rvs with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$, is given by

  $$E[h(X,Y)] = \sum_x \sum_y h(x,y) \cdot p(x,y) \qquad \text{(if } X \text{ and } Y \text{ are discrete)}$$

  $$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y)\, dx\, dy \qquad \text{(if } X \text{ and } Y \text{ are continuous)}$$

  This is sometimes referred to as the *Law of the Unconscious Statistician.*

- Property: (Linearity of Expectation)

  Let X and Y be random variables. Then, for any functions h_1, h_2 and any constants a_1, a_2, b,

  $$E[a_1 h_1(X,Y) + a_2 h_2(X,Y) + b] = a_1 E[h_1(X,Y)] + a_2 E[h_2(X,Y)] + b$$

- Theorem

  Let X and Y be *independent* random variables. If $h(X, Y) = g_1(X) \cdot g_2(Y)$, then

  $$E[h(X,Y)] = E[g_1(X) \cdot g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)]$$

4.2.2 Covariance

A measurement of how strongly two random variables are related to each other.

- Definition

  The **covariance** between two rvs X and Y is

  $$\begin{aligned} \operatorname{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) && \text{(if } X \text{ and } Y \text{ are discrete)} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\, dx\, dy && \text{(if } X \text{ and } Y \text{ are continuous)} \end{aligned}$$

- Proposition

  For any two random variables X and Y,

  1. $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
  2. $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$
  3. (Covarance shortcut formula) $\operatorname{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y$
  4. (Distributive property of covariance) For any rv Z and any constants, a, b, c,

  $$\operatorname{Cov}(aX + bY + c, Z) = a \operatorname{Cov}(X, Z) + b \operatorname{Cov}(Y, Z)$$

4.2.3 Correlation

Measures how strongly two random variables are related to each other without the influence of unit.

- Definition

  The **correlation coefficient** of two rvs X and Y, denoted by $\text{Corr}(X, Y)$, or $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- Proposition

  For any two rvs X and Y,

  1. $\text{Corr}(X, Y) = \text{Corr}(Y, X)$
  2. $\text{Corr}(X, X) = 1$
  3. (Scale invariance property) If a, b, c, d are constants and $ac > 0$, then

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

  4. $-1 \leq \text{Corr}(X, Y) \leq 1$

- Proposition

  1. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.
  2. $\rho = 1$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.
  3. Two rvs X and Y are uncorrelated iff $E[XY] = \mu_X \cdot \mu_Y$.

     Notice:

     If X and Y are independent, then $E[g_1(X)g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)]$.
     If X and Y are uncorrelated, then $E[XY] = \mu_X \cdot \mu_Y$.
     Thus independence is stronger than zero correlation, the latter one being the special case corresponding to $g_1(X) = X$ and $g_2(Y) = Y$.

- Comments:

  1. The sign of ρ indicates whether X and Y are positively or negatively related, and the magnitude of ρ describes the strength of that relationship on an absolute 0–1 scale.
  2. A value of ρ near 1 does not necessarily imply that increasing the value of X *causes* Y to increase. It implies only that large X values are associated with large Y values. In other words association (a high correlation) is not the same as causation.

4.3 Properties of Linear Combinations

4.3.1 Linear Combinations of Random Variables

- Definition

  A linear combination of random variables refers to anything of the form

  $$a_1X_1 + \cdots + a_nX_n + b$$

  where the X_is are random variables and the a_is and b are numerical constants.

- Theorem

  Let the rvs $X_1, X_2, \cdots, X_n$ have mean values $\mu_1, ..., \mu_n$ and standard deviations $\sigma_1, ..., \sigma_n$, respectively.

  1. Whether or not the X_is are independent,

     $$\begin{aligned} E(a_1X_1 + \cdots + a_nX_n + b) &= a_1E(X_1) + \cdots + a_nE(X_n) + b \\ &= a_1\mu_1 + \cdots + a_n\mu_n + b \end{aligned}$$

     and

     $$\begin{aligned} \text{Var}(a_1X_1 + \cdots + a_nX_n + b) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum\sum_{i<j} a_ia_j\text{Cov}(X_i, X_j) \end{aligned}$$

  2. If $X_1, \cdots, X_n$ are independent

     $$\begin{aligned} \text{Var}(a_1X_1 + \cdots + a_nX_n + b) &= a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n) \\ &= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \end{aligned}$$

     and

     $$\text{SD}(a_1X_1 + \cdots + a_nX_n + b) = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2}$$

- Corollary

  For two rvs X_1 and X_2, and any constants a_1, a_2, b,

  $$E(a_1X_1 + a_2X_2 + b) = a_1E(X_1) + a_2E(X_2) + b$$

  and

  $$\text{Var}(a_1X_1 + a_2X_2 + b) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + 2a_1a_2\text{Cov}(X_1, X_2)$$

  In particular, if X_1 and X_2 are independent,

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

  which can also be written as

  $$[\text{SD}(X_1 + X_2)]^2 = [\text{SD}(X_1 - X_2)]^2 = \text{SD}(X_1)^2 + \text{SD}(X_2)^2$$

  statisticians sometimes call this property the *Pythagorean Theorem.*

- Definition

  Let X and Y be independent, continuous rvs with marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively. Then the pdf of the rv $W = X + Y$ is given by

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

  [In mathematics, this integral operation is known as the **convolution** of $f_X(x)$ and $f_Y(y)$ and is sometimes denoted $f_W = f_X \star f_Y$.] The limits of integration are determined by which x values make both $f_X(x) > 0$ and $f_Y(w - x) > 0$.

4.3.2 Moment Generating Functions for Linear Combinations

- Proposition

  Let $X_1, \cdots, X_n$ be independent random variables with moment generating functions $M_{X_1}(t), \cdots, M_{X_n}(t)$, respectively. Then the MGF of the linear combination $Y = a_1X_1 + \cdots + a_nX_n + b$ is

  $$M_Y(t) = e^{bt}M_{X_1}(a_1t)\cdots M_{X_n}(a_nt)$$

  In the special case that $a_1 = \cdots = a_n = 1$ and $b = 0$, so $Y = X_1 + \cdots + X_n$,

  $$M_Y(t) = M_{X_1}(t)\cdots M_{X_n}(t)$$

  That is, the MGF of a sum of independent rvs is the product of the individual MGFs.

- Proposition (Linear combination of independent Normally Distributed rvs)

  If $X_1, X_2, ..., X_n$ are independent, normally distributed rvs (with possibly different means and/or sds), then any linear combination of the X_is also has a normal distribution. In particular, the sum of independent normally distributed rvs itself has a normal distribution, and the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

- Proposition (Linear combination of independent Poisson rvs)

  Suppose $X_1, X_2, ..., X_n$ are independent Poisson random variables, where X_i has mean μ_i. Then $Y = X_1 + \cdots + X_n$ also has a Poisson distribution, with mean $\mu_1 + \cdots + \mu_n$.

- Proposition (Linear combination of independent Exponential rvs)

  Suppose $X_1, \cdots, X_n$ are independent exponential random variables with common parameter λ. Then $Y = X_1 + \cdots + X_n$ has a gamma distribution, with parameters $\alpha = n$ and $\beta = \frac{1}{\lambda}$ (aka the Erlang distribution).

4.4 Conditional Probability Distributions, Expectation and Variance

4.4.1 Conditional Probability Distributions

- Definition (discrete case)

  Let X and Y be two discrete random variables with joint pmf $p(x, y)$ and marginal X pmf $p_X(x)$. Then for any x value such that $p_X(x) > 0$, the **conditional probability mass function of Y given $X = x$** is

  $$p_{Y|X}(y \,|\, x) = \frac{p(x, y)}{p_X(x)}$$

- Definition (continuous case)

  Let X and Y be two continuous random variables with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$. Then for any x value such that $f_X(x) > 0$, the **conditional probability density function of Y given $X = x$** is

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

- Proposition

  Let X and Y be two random variables, then X and Y are independent iff

  $$p_{Y|X}(y \,|\, x) = p_Y(y) \qquad \text{(if } X \text{ and } Y \text{ are discrete)}$$
  $$f_{Y|X}(y \,|\, x) = f_Y(y) \qquad \text{(if } X \text{ and } Y \text{ are continuous)}$$

4.4.2 Conditional Expectation

- Conditional Expectation (discrete case)

  Let X and Y be two discrete random variables with conditional probability mass function $p_{Y|X}(y \,|\, x)$. Then the **conditional expectation (or conditional mean) of Y given $X = x$** is

  $$\mu_{Y|X=x} = E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y \,|\, x)$$

  More generally, the conditional mean of any function $h(Y)$ is given by

  $$E(h(Y)|X = x) = \sum_y h(y) \cdot p_{Y|X}(y \,|\, x)$$

- Conditional Expectation (continuous case)

  Let X and Y be two continuous random variables with conditional probability density function $f_{Y|X}(y \,|\, x)$. Then the **conditional expectation (or conditional mean) of Y given $X = x$** is

  $$\mu_{Y|X=x} = E(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \,|\, x)\, dy$$

  More generally, the conditional mean of any function $h(Y)$ is given by

  $$E(h(Y)|X = x) = \int_{-\infty}^{\infty} h(y) \cdot f_{Y|X}(y \,|\, x)\, dy$$

4.4.3 Conditional Variance

Let X and Y be two random variables, then the **conditional variance of Y given $X = x$** is

$$\sigma^2_{Y|X=x} = \mathrm{Var}(Y|X = x) = E[(Y - \mu_{Y|X=x})^2 \,|\, X = x] = E(Y^2 \,|\, X = x) - \mu^2_{Y|X=x}$$

4.4.4 The Laws of Total Expectation and Variance

- Law of Total Expectation

  For any two random variables X and Y,

  $$E[E(Y \mid X)] = E(Y)$$

  (This is sometimes referred to as computing $E(Y)$ by means of *iterated expectation.*)

  Remark:
  $E(Y)$ is a weighted average of the conditional means $E(Y|X=x)$, where the weights are given by the pmf or pdf of X.

- Law of Total Variance

  For any two random variables X and Y,

  $$\mathrm{Var}(Y) = \mathrm{Var}[E(Y|X)] + E[\mathrm{Var}(Y|X)]$$

4.5 Limit Theorems

4.5.1 Random Samples

- Definition

  The rvs $X_1, X_2, \cdots, X_n$ are said to be **independent and identically distributed (iid)** if

  1. The X_is are independent rvs.
  2. Every X_i has the same probability distribution.

  Such a collection of rvs is also called a (simple) **random sample** of size n.

  Remark:
  If sampling is either with replacement or from a (potentially) infinite population, Conditions 1 and 2 are satisfied exactly.
  These conditions will be approximately satisfied if sampling is without replacement, yet the sample size n is much smaller than the population size N. In practice, if $n/N \leq .05$ (at most 5% of the population is sampled), we proceed as if the X_is form a random sample.

- Definition

  Let $X_1, X_2, \cdots, X_n$ be a random sample of size n, then

  $$T = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{T}{n}$$

  the rv T is called the **sample total**, and the rv $\overline{X}$ is called the **sample mean**.

- Proposition

  Suppose $X_1, \cdots, X_n$ are iid with common mean μ and common standard deviation σ. T and $\overline{X}$ have the following properties:

  1. $E(T) = n\mu$
  2. $\text{Var}(T) = n\sigma^2$ and $\text{SD}(T) = \sqrt{n}\sigma$
  3. If the X_is are normally distributed, then T is also normally distributed.

  1. $E(\overline{X}) = \mu$
  2. $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$ and $\text{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$
  3. If the X_is are normally distributed, then $\overline{X}$ is also normally distributed.

4.5.2 The Central Limit Theorem

Let $X_1, \cdots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then, in the limit as $n \to \infty$, the standardized versions of T and $\overline{X}$ have the standard normal distribution. That is,

$$\lim_{n\to\infty} P(\frac{T - n\mu}{\sqrt{n}\sigma} \leq z) = P(Z \leq z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z) = P(Z \leq z) = \Phi(z)$$

where Z is a standard normal rv. It is customary to say that T and $\overline{X}$ are **asymptotically normal**.

Thus when n is sufficiently large, the sample total T has approximately a normal distribution with mean $\mu_T = n\mu$ and standard deviation $\sigma_T = \sqrt{n}\sigma$.

Equivalently, for large n the sample mean $\overline{X}$ has approximately a normal distribution with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Remark
The accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled. If the underlying distribution is symmetric and there is not much probability far out in the tails, then the approximation will be good even for a small n, whereas if it is highly skewed or has "heavy" tails, then a large n will be required.

4.5.3 Other Applications of the Central Limit Theorem

- Corollary

  Consider an event A in the sample space of some experiment with $p = P(A)$. Let $X =$ the number of times A occurs when the experiment is repeated n independent times, and define

  $$\hat{P} = \hat{P}(A) = \frac{X}{n}$$

  Then

  1. $\mu_{\hat{P}} = E(\hat{P}) = p$
  2. $\sigma_{\hat{P}} = \mathrm{SD}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$
  3. As n increases, the distribution of $\hat{P}$ approaches a normal distribution.

  In practice, Property 3 is taken to say that $\hat{P}$ is approximately normal, provided that $np \geq 10$ and $n(1-p) \geq 10$.

- Corollary

  CLI justifies normal approximations to the following distributions:

  - Poisson, when μ is large.
  - Negative binomial, when r is large.
  - Gamma, when α is large.

- Proposition

  Let $X_1, \cdots, X_n$ be a random sample from a distribution for which only positive values are possible $[P(X_i > 0) = 1]$. Then if n is sufficiently large, the product $Y = X_1 X_2 \cdots X_n$ has approximately a lognormal distribution; that is, $\ln(Y)$ has approximately a normal distribution.

4.5.4 The Law of Large Numbers

- LAW OF LARGE NUMBERS

  If $X_1, \cdots, X_n$ is a random sample from a distribution with mean μ and finite variance, then $\overline{X}$ converges to μ

  1. In mean square: $E[(\overline{X} - \mu)^2] \to 0$ as $n \to \infty$
  2. In probability: $P(|\overline{X} - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$

- LAW OF LARGE NUMBERS (Proportions)

  If the X_i are iid Bernoulli(p) rvs, then the sample proportion $\hat{P}$ converges to the "true" proportion p

  1. In mean square: $E[(\hat{P} - p)^2] \to 0$ as $n \to \infty$
  2. In probability: $P(|\hat{P} - p| \geq \varepsilon) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$

- Summary

  In statistical language, the Law of Large Numbers states that $\overline{X}$ is a consistent estimator of μ, and $\hat{P}$ is a consistent estimator of p.

4.6 Transformations of Jointly Distributed Random Variables

4.6.1 The Joint Distribution of Two New Random Variables

- Notations

  Let X_1 and X_2 be two continuous rvs, consider forming two new random variables

$$Y_1 = u_1(X_1, X_2) \qquad Y_2 = u_2(X_1, X_2)$$

  Let

$$\begin{aligned} f(x_1, x_2) &= \text{the joint pdf of the two original variables} \\ g(y_1, y_2) &= \text{the joint pdf of the two new variables} \end{aligned}$$

  The $u_1(\cdot)$ and $u_2(\cdot)$ functions express the new variables in terms of the original ones. The general result presumes that these functions can be inverted to solve for the original variables in terms of the new ones:

$$X_1 = v_1(Y_1, Y_2) \qquad X_2 = v_2(Y_1, Y_2)$$

  Then let

$$S = \{(x_1, x_2) : f(x_1, x_2) > 0\} \qquad T = \{(y_1, y_2) : g(y_1, y_2) > 0\}$$

  That is, S is the region of positive density for the original variables and T is the region of positive density for the new variables; T is the "image" of S under the transformation.

- TRANSFORMATION THEOREM (bivariate case)

  Suppose that the partial derivative of each $v_i(y_1, y_2)$ with respect to both y_1 and y_2 exists and is continuous for every $(y_1, y_2) \in T$. Form the 2×2 matrix

$$M = \begin{pmatrix} \frac{\partial v_1(y_1,y_2)}{\partial y_1} & \frac{\partial v_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial v_2(y_1,y_2)}{\partial y_1} & \frac{\partial v_2(y_1,y_2)}{\partial y_2} \end{pmatrix}$$

  The determinant of this matrix, called the *Jacobian*, is

$$\det(M) = \frac{\partial v_1}{\partial y_1} \cdot \frac{\partial v_2}{\partial y_2} - \frac{\partial v_1}{\partial y_2} \cdot \frac{\partial v_2}{\partial y_1}$$

  The joint pdf for the new variables then results from taking the joint pdf $f(x_1, x_2)$ for the original variables, replacing x_1 and x_2 by their expressions in terms of y_1 and y_2, and finally multiplying this by the absolute value of the Jacobian:

$$g(y_1, y_2) = f(v_1(y_1, y_2), v_2(y_1, y_2)) \cdot |\det(M)| \qquad (y_1, y_2) \in T$$

4.6.2 The Joint Distribution of More Than Two New Variables

Suppose that the partial derivative of each $v_i(y_1, \cdots, y_n)$ with respect to all y_i exists and is continuous for every $(y_1, \cdots, y_n) \in T$. Form the $n \times n$ matrix

$$M = \begin{pmatrix} \frac{\partial v_1}{\partial y_1} & \cdots & \frac{\partial v_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial v_n}{\partial y_1} & \cdots & \frac{\partial v_n}{\partial y_n} \end{pmatrix}$$

The joint pdf for the new variables then results from taking the joint pdf $f(x_1, \cdots, x_n)$ for the original variables, replacing $x_1, \cdots, x_n$ by their expressions in terms of $y_1, \cdots, y_n$, and finally multiplying this by the absolute value of the Jacobian:

$$g(y_1, \cdots, y_n) = f(v_1, \cdots, v_n) \cdot |\det(M)| \qquad (y_1, \cdots, y_n) \in T$$