National University of Singapore
Department of Statistics and Data Science

ST2334: Probability and Statistics

Formulae and Facts

1. **Probability Rules**
For events A and B in the sample space S:

(i) $P(A) = P(A \cap B) + P(A \cap B')$

(ii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(iii) $P(B|A) = \frac{P(A\cap B)}{P(A)}$

(iv) $P(A \cap B) = P(A)P(B|A)$

(v) $P(B) = P(A)P(B|A) + P(A')P(B|A')$

(vi) $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

(vii) $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A')P(B|A')}$

Note: We need $P(A) > 0$ for items (1iii) onwards, and $P(B) > 0$ for items (1vi) onwards.

2. **Random Variables**

(i) Let X and Y be two random variables, and let a and b be any real numbers. Then

$$E(aX+b) = aE(X)+b$$
$$E(X+Y) = E(X)+E(Y)$$

(ii) Let $g(\cdot)$ be an arbitrary function. Then

$$E[g(X)] = \begin{cases} \sum_{x\in R_X} g(x)f(x), & \text{if } X \text{ is discrete;} \\ \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous.} \end{cases}$$

3. **Joint Distributions**

(i) Random variables X and Y are independent if and only if for **any** x and y,

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

(ii) Consider any two variable function $g(x,y)$. Then

$$E[g(X,Y)] = \begin{cases} \sum_x\sum_y g(x,y)f_{X,Y}(x,y), & \text{if } (X,Y) \text{ discrete;} \\ \iint_{\mathbb{R}^2} g(x,y)f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x, & \text{if } (X,Y) \text{ continuous.} \end{cases}$$

(iii) The covariance of X and Y is defined to be

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

(iv) $\mathrm{cov}(aX+b, cY+d) = ac\cdot\mathrm{cov}(X,Y)$

(v) $V(aX+bY) = a^2V(X)+b^2V(Y)+2ab\,\mathrm{cov}(X,Y)$

4. **Common Probability Distributions**

(i) Binomial(n,p)

$$f_X(x) = \binom{n}{x}p^x(1-p)^{n-x}, \quad x = 0,1,2,\ldots,n.$$

$E(X) = np$, $V(X) = np(1-p)$.

(ii) Negative Binomial(k,p)

$$f_X(x) = \binom{x-1}{k-1}p^k(1-p)^{x-k}, \quad x = k,k+1,\ldots.$$

$E(X) = \frac{k}{p}$, $V(X) = \frac{(1-p)k}{p^2}$.

(iii) Geometric(p)

$$f_X(x) = (1-p)^{x-1}p, \quad x = 1,2,\ldots.$$

$E(X) = \frac{1}{p}$, $V(X) = \frac{1-p}{p^2}$.

(iv) Poisson(λ)

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2,\ldots.$$

$E(X) = \lambda$, $V(X) = \lambda$.

(v) Uniform(a,b)

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b; \\ 0, & \text{otherwise.} \end{cases}$$

$E(X) = \frac{a+b}{2}$, $V(X) = \frac{(b-a)^2}{12}$.

(vi) Exponential(λ)

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \ge 0; \\ 0, & \text{if } x < 0. \end{cases}$$

$E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$.

(vii) Normal(μ,σ^2)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

$E(X) = \mu$, $V(X) = \sigma^2$.

5. **Confidence Intervals / Test Statistics: Population Mean**
The table below gives

- the $100(1-\alpha)\%$ confidence interval formulas for the population mean μ,
- the test statistics for the (null) hypothesis: $H_0 : \mu = \mu_0$.

Case	Population	σ	n	Confidence Interval	Test Statistic
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$Z = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim N(0,1)$

Note that n is considered large when $n \geq 30$.

6. **Confidence Intervals / Test Statistics: Difference of Population Means**
The table below, **for two independent samples**, gives

- the $100(1-\alpha)\%$ confidence interval formulas for $\mu_1 - \mu_2$,
- the test statistics for the (null) hypothesis: $H_0 : \mu_1 = \mu_2$.

Populations	σ_1, σ_2	n_1, n_2	Confidence Interval	Test Statistic
Any	known, unequal	$n_1 \geq 30, n_2 \geq 30$	$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
Normal	known, unequal	any	$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
Any	unknown, unequal	$n_1 \geq 30, n_2 \geq 30$	$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N(0,1)$
Normal	unknown, equal	$n_1 < 30, n_2 < 30$	$(\bar{x}-\bar{y}) \pm t_{n_1+n_2-2,\alpha/2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$T = \frac{(\bar{X}-\bar{Y})}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$
Any	unknown, equal	$n_1 \geq 30, n_2 \geq 30$	$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$Z = \frac{(\bar{X}-\bar{Y})}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$

Here $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ is the pooled sample variance.
For dependent samples, consider the sample $D_i = X_i - Y_i$, and use the results in Section 5.

7. **Miscellaneous**

(i) **Roots of the quadratic equation**
For the equation $ax^2+bx+c=0$, $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$.

(ii) **Sum of the first n terms of a geometric series**
$a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a(1-r^n)}{1-r}$.