Formulas

Probability

Definitions (Noteworthy ones)
- Sure Event: sample space
- Sample space depends on problem of interest

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

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

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

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

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

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

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

Independence

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

$P(A) = P(A|B)$, $P(B) \neq 0$

If $P(A) > 0$ and $P(B) > 0$, then $A \perp B \Rightarrow A$ and B not mutually exclusive

Contrapositive: , then A and B mutually exclusive $\Rightarrow A \not\perp B$

$A \perp B \Rightarrow A \perp B', A' \perp B, A' \perp B'$

Mutually exclusive

$P(A \cup B) = P(A) + P(B)$

$P(A \cap B) = \varnothing$

Expectation, Variance, Covariance

- $E(X) = \sum_{R_X} xf(x)$ or $\int_{R_X} xf(x)$
- $E(g(X)) = \sum_{R_X} g(x)f(x)$ or $\int_{R_X} g(x)f(x)$
- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$

- $E(X_1 + X_2 + ... + X_n) = nE(X)$
- $X \perp Y \implies E(XY) = E(X)E(Y)$ **INVERSE IS NOT NECESSARILY TRUE!!!**

- $V(X) = E[(X - \mu_x)^2]$
- $V(X) = E(X^2) - [E(X)]^2$
- $V(aX + b) = a^2V(X)$
- $V(aX_1 + bX_2) = a^2V(X) + b^2V(X)$, X_1, X_2 are random **observations** of X

- $cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
  $= E(XY) - E(X)E(Y)$
  $= \int\int_{R_{X,Y}} (x - \mu_x)(y - \mu_y)f_{X,Y}(x, y)\, dxdy$
- $X \perp Y \implies cov(X, Y) = 0$ **INVERSE IS NOT NECESSARILY TRUE!!!**
- $cov(aX + b, cY + d) = ac \cdot cov(X, Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \cdot cov(X, Y)$
- $V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n) + 2\sum_{j>i} cov(X_i, X_j)$

Random Variables

> **Definitions (Noteworthy ones)**
> - **Random Variable**: A function which assigns a real number to every $s \in S$

Discrete

Probablity Function

(1) $f(x_i) \geq 0$ for all $x_i \in R_X$;

(2) $f(x) = 0$ for all $x \notin R_X$;

(3) $\sum_{i=1}^{\infty} f(x_i) = 1$, or $\sum_{x_i \in R_X} f(x_i) = 1$.

Cumulative Distribution Function

$F(x) = P(X \leq x)$

- $P(a \leq X \leq b) = F(b) - F(a-)$
- $P(a < X < b) = F(b-) - F(a)$

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 0.3, & 1 \le x < 3 \quad \text{F(1) = P(x ≤ 1) = F(2)} \\ 0.4, & 3 \le x < 4 \quad \text{F(3) = P(x ≤ 3)} \\ 0.45, & 4 \le x < 6 \quad \text{F(4) = P(x ≤ 4) = F(5)} \\ 0.6, & 6 \le x < 12 \quad \text{F(6) = P(x ≤ 6) = F(6) = F(7) = ... = F(11)} \\ 1, & 12 \le x \quad \text{F(12) = P(x ≤ 12)} \end{cases}$$

x	1	3	4	6	12
$f_X(x)$	0.3	0.1	0.05	0.15	0.4

Continuous

Probability Density Function

(1) $f(x) \ge 0$ for all $x \in R_X$; and $f(x) = 0$ for $x \notin R_X$.

(2) $\int_{R_X} f(x)dx = 1$.

(3) For any a and b such that $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)dx.$$

🚨 $f_X(x)$ can be > 1

For any arbitrary specific value x_0, we have

$$P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0.$$

This gives an example of "$P(A) = 0$, but A is not necessarily $\emptyset$."

Cumulative Distribution Function

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

$$f(x) = \frac{dF(x)}{dx}.$$

$F(x)$ for $a \le X \le b$ = $\int_a^x f(t)\,dx + F(a)$ for piecewise functions

2D-RVs

Marginal

$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$

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

Conditional

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

$f_{X|Y}(x|y)$ is a **p.f.** for X

$P(X \le x|Y=y) = \int_{-\infty}^{x} f_{X|Y}(x,y)dx$

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

$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = 1$, but $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dy$ need not = 1

Independence

$f_{X,Y}(x,y) = f_X(x)f_Y(y),\ \forall (x,y) \in R_{X,Y}$

1. $P(X \le x, Y \le y) = P(X \le x)P(Y \le y) = F(x)F(y) = F_{X,Y}(x,y)$
2. $g_1(X)$ and $g_2(Y)$ are independent for any arbitrary $g_1(\cdot)$ and $g_2(\cdot)$
   a. Consequently, $E(XY) = E(X)E(Y); E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)]$
3. If $f_X(x) > 0$, then $f_{Y|X}(y,x) = f_Y(y)$

   If $f_Y(y) > 0$, then $f_{X|Y}(x,y) = f_X(x)$

Conditions:

1. $f_{X,Y}(x,y)$ can be **factorised to the form:** $c \cdot g_1(x)g_2(y)$
2. **Range of X does not depend on Y** and vice versa

- If $f_{Y|X}(y|x)$ contains x in the formula, then X and Y are not independent, and vice versa

Sampling

Unbiased estimator: mean value equals to true value of parameter i.e. $E(\hat{\Theta}) = \theta$

THEOREM 1 (CENTRAL LIMIT THEOREM (CLT))
If $\overline{X}$ is the mean of a random sample of size n taken from a population having mean μ and finite variance σ^2, then, as $n \to \infty$,

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1),$$

or equivalently

$$\overline{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

WHAT IS THE BIG DEAL?
The Central Limit Theorem states that, under rather general conditions, for large n, sums and means of random samples drawn from a population follows the normal distribution closely.

Note that if the random sample comes from a normal population, $\overline{X}$ is normally distributed regardless of the value of n.

THEOREM 7 (LAW OF LARGE NUMBERS (LLN))
If $X_1, \ldots, X_n$ are independent random variables with the same mean μ and variance σ^2, then for any $\varepsilon \in \mathbb{R}$,

$$P(|\overline{X}-\mu| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Sample Mean of X

$$\mu_X = E(X)$$

Sample Variance of X (unbiased estimator)

$$\begin{aligned} S^2 = \sigma_X^2 &= \frac{\sum(X-\overline{X})^2}{n-1} \\ &= \frac{1}{n-1}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right) \\ &= \frac{1}{n-1}\left(\sum(X^2) - n\overline{X}^2\right) \end{aligned}$$

Sample Mean of $\overline{X}$

$$\mu_{\overline{X}} = \mu_X$$

Sample Variance of $\overline{X}$

$$\sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n}$$

Central Limit Theorem

Distribution (X)	Variance	Size	Distribution ($\overline{X}$)
Normal	Known	Any	Standard Normal
Any	Known	Large	Standard Normal
Any	Unknown	Large	Standard Normal
Normal	Unknown	Small	t Distribution

Approximations

- $X \sim B(n,p) \approx X \sim \text{Poisson}(np)$,
  - $n \geq 20$ and $p \leq 0.05$, or $n \geq 100$ and $np \leq 10$
- $X \sim B(n,p) \approx X \sim N(np, npq)$,
  - $np > 5$ and $n(1-p) > 5$

In this example, we have made the **continuity correction** to improve the approximation. In general, we have

(a) $P(X = k) \approx P(k - 1/2 < X < k + 1/2)$;

(b) $P(a \leq X \leq b) \approx P(a - 1/2 < X < b + 1/2)$;
$P(a < X \leq b) \approx P(a + 1/2 < X < b + 1/2)$;
$P(a \leq X < b) \approx P(a - 1/2 < X < b - 1/2)$;
$P(a < X < b) \approx P(a + 1/2 < X < b - 1/2)$.

(c) $P(X \leq c) = P(0 \leq X \leq c) \approx P(-1/2 < X < c + 1/2)$.

(d) $P(X > c) = P(c < X \leq n) \approx P(c + 1/2 < X < n + 1/2)$.

Distributions

	$P(X=x)$ / $f_X(x)$	$P(X \leq x)$ / $F_X(x)$	$E(X)$	$Var(X)$	Misc
Discrete Uniform (D)	$1/k$		$\frac{1}{k}\sum_{i=1}^{k} x_i$	$\frac{1}{k}\sum x^2 - E(X)$	
Bernoulli	$p^x(1-p)^{1-x}$		p	$p(1-p)$	
Binomial (D)	$\binom{n}{x}p^x(1-p)^{n-x}$		np	$np(1-p)$	$X+Y \sim B(2n,p)$, if $X \perp Y$
Negative Binomial (D)	$\binom{x-1}{k-1}p^k(1-p)^{x-k}$		k/p	$(1-p)k/p^2$	
Poisson (D)	$(e^{-\lambda}\lambda^k)/k!$, $\lambda > 0$		λ	λ	$X+Y \sim Poisson(\lambda_x + \lambda_y)$
Geometric (D)	$(1-p)^{x-1}p$		$1/p$	$(1-p)/p^2$	$P(X > x) = (1-p)^x$ $P(X = x_2 \mid X > x_1) = P(X = x_2 - x_1)$
Continuous Uniform (C)	$1/(b-a)$ $a \leq x \leq b$	$(x-a)/(b-a)$ $a \leq x \leq b$	$(a+b)/2$	$(b-a)^2/12$	
Exponential (C)	$\lambda e^{-\lambda x}, \lambda > 0, x > 0$	$1-e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	$P(X > x_2 \mid X > x_1) = P(X > x_2 - x_1)$
Normal (C)	$\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$	$P(Z \leq \frac{X-\mu}{\sigma})$	μ	σ^2	$aX+b \sim N(a\mu+b, a^2\sigma^2)$ $aX \pm bY \sim N(a\mu_1 \pm b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$, if $X \perp Y$ Range: $\mu \pm s.d(\sigma)$

Special Distributions

$Y \sim \chi^2(n)$	Definition: $Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ $E(Y) = n, V(Y) = 2n$ $Y \approx N(n, 2n), n \to \infty$ $X \perp Y \implies X + Y \sim \chi^2(m+n)$ $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, if normal
$T \sim t(n)$	Definition: $T = \frac{Z}{\sqrt{U/n}}$, if $Z \perp U$ and $U \sim \chi^2(n)$ $E(T) = 0, V(T) = n/(n-2)$ $T \approx N(0,1), n \geq 30$ $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
$X \sim F(m, n)$	Definition: $F = \frac{U/m}{V/n}, U \sim \chi^2(m) \perp V \sim \chi^2(n)$ $E(X) = n/(n-2), V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $X \sim F(m,n) \implies 1/X \sim F(n.m)$ $F(n, m; \alpha/2) = 1/F(m, n; 1-\alpha/2)$ $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$, given two independent populations

Confidence Intervals and MaxEoE

Max Error of Estimate: $100\alpha\%$ probability that $|\overline{X} - \mu|$ is less than E
Confidence Interval: If many CIs are taken, about $(1-\alpha)\%$ of them will contain the true parameter μ
INCORRECT DEFINITION: "The probability that μ is contained in the CI is $(1-\alpha)\%$"

DIFFERENT CASES:

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
II	any	known	large	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2} \cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot s}{E_0}\right)^2$

CONFIDENCE INTERVALS FOR THE MEAN:
The table below gives the $(1-\alpha)$ confidence interval (formulas) for the population mean.

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

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

Sample size	Variance	Formula for $(1-\alpha)$ confidence interval
Normal or large	Known + Unequal	$(\overline{x}-\overline{y}) \pm z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
Large	Unknown + Unequal	$(\overline{x}-\overline{y}) \pm z_{a/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
Normal + Small	Unknown + Equal	$(\overline{x}-\overline{y}) \pm$ $(t_{n_1+n_2-2;\alpha/2})S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
Large	Unknown + Equal	$(\overline{x}-\overline{y}) \pm (z_{\alpha/2})S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

Sample size	Formula for $(1-\alpha)$ confidence interval
Normal + Small	$\overline{d} \pm (t_{n-1;\alpha/2})\frac{s_D}{\sqrt{n}}$
Large	$\overline{d} \pm (z_{\alpha/2})\frac{s_D}{\sqrt{n}}$

Hypothesis Testing

Null Hypothesis: Default assumption. Want to show is false
Alternative Hypothesis: Something we want to prove
Type I error: Reject H_0 when H_0 is true
Type II error: Not rejecting H_0 when H_0 is false
α: Level of significance = $P(\text{Reject } H_0 | H_0 \text{ is true})$
β: $P(\text{Do not reject } H_0 | H_0 \text{ is false})$
1 - β: Power = $P(\text{Reject } H_0 | H_0 \text{ is false})$
p-value: Probability of obtaining a test statistic at least as extreme as the observed sample value, given H_0 is true (observed level of significance)

Population	Variance	n	Statistic	Rejection Criteria: 1-tailed	Rejection Criteria: 2-tailed
Normal	Known	Any	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$Z < -z_\alpha$ or $Z > z_\alpha$ $P(Z < -z_{calc}) < \alpha$ or $P(Z > z_{calc}) < \alpha$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ $2P(Z > z_{calc}) < \alpha$
Any	Known	Large	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$Z < -z_\alpha$ or $Z > z_\alpha$ $P(Z < z_{calc}) < \alpha$ or $P(Z > z_{calc}) < \alpha$	$Z < -z_{\alpha/2}$ or$Z > z_{\alpha/2}$ $2P(Z > z_{calc}) < \alpha$
Normal	Unknown	Small	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$	$t < -t_{n-1,\alpha}$ or $t > t_{n-1,\alpha}$	$t < -t_{n-1,\alpha/2}$ or $t > t_{n-1,\alpha/2}$
Any	Unknown	Large	$Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$Z < -z_\alpha$ or $Z > z_\alpha$ $P(Z < z_{calc}) < \alpha$ or $P(Z > z_{calc}) < \alpha$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ $2P(Z > z_{calc}) < \alpha$
Both Normal and independent	Both known	Both Large, if not normal	$Z = \frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$, $\delta_0 = \overline{D}$ (population) $D_i = X_i - Y_i$	Same as blue	Same as blue
Any	Both unknown	Large	$Z = \frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	Same as blue	Same as blue
Normal	Both unknown but equal	Small	$Z = \frac{\overline{X}-\overline{Y}-\delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$	Same as blue	Same as blue
Paired	Unknown	Small	$T = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$	Same as red	Same as red
Paired	Unknown	Large	$T = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}} \approx Z \sim N(0,1)$	Same as blue	Same as blue

- Statements will always be about the **means** of a **population**