# Formulas That May Be Needed

## 1 Laws of Probability

- If $A$ and $B$ are mutually exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$.
- If $A$ and $B$ are independent events, then
  - $P(A \text{ and } B) = P(A) \times P(B)$,
  - $P(A \mid B) = P(A)$.
- If $P(A \text{ and } B) = P(A) \times P(B)$ or $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$, then
  - $A$ and $B$ are independent events.
- If $A$ and $B$ are two events and $P(B) \neq 0$, the conditional probability of $A$ given $B$ is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B \mid A) \times P(A)}{P(B)} .$$

## 2 Discrete Random Variables (RV from now on)

$$\bar{X} = E(X) = \mu_X = \sum_{i=1}^{n} P(X = x_i)\, x_i \qquad \mathrm{VAR}(X) = \sigma_X^2 = \sum_{i=1}^{n} P(X = x_i)(x_i - \mu_X)^2$$

$$\text{Std Dev}(X) = \sigma_X = \sqrt{\mathrm{VAR}(X)}$$

## 3 Binomial Distribution with Parameters $n$ and $p$

$$\mu_X = np \qquad \sigma_X^2 = np(1-p) \qquad P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

## 4 Two Random Variables

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{n} P(X = x_i \text{ and } Y = y_i)(x_i - \mu_X)(y_i - \mu_Y)$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

If $X$ and $Y$ are independent, then $\mathrm{Cov}(X, Y) = 0$ and $\mathrm{Corr}(X, Y) = 0$.

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

$$\begin{aligned} \mathrm{VAR}(aX + bY + c) &= a^2 \mathrm{VAR}(X) + b^2 \mathrm{VAR}(Y) + 2ab\, \mathrm{Cov}(X, Y) \\ &= a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab\, \sigma_X \sigma_Y \, \mathrm{Corr}(X, Y) \end{aligned}$$

## 5 Uniform Distribution between $a$ and $b$

$$E(X) = \frac{a+b}{2} \qquad \text{VAR}(X) = \frac{(b-a)^2}{12} \qquad P(X \le x) = \frac{x-a}{b-a} \quad \text{if } a \le x \le b$$

## 6 Normal Distribution

- If $X$ is a normal distribution with mean $\mu$ and standard deviation $\sigma$, then $P(X \le x) = F(\frac{x-\mu}{\sigma})$, where $F(z)$ can be read from the "normal" table and $z = \frac{x-\mu}{\sigma}$.
- If $X$ and $Y$ are Normally distributed, then so is $aX + bY + c$.
- Assume that $X_1, .., X_n$ are independent and identically distributed, $E(X_i) = \mu$, and $\text{VAR}(X_i) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$ be the sum and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the average, then:
  - **Central Limit Theorem for the sum**. If $n$ is moderately large (say, 30 or more) then $S_n$ is approximately Normally distributed with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.
  - **Central Limit Theorem for the sample mean**. If $n$ is moderately large, then $\bar{X}$ is approximately Normally distributed with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- A binomial distribution can be approximated with a normal (with the correct parameters $\mu$ and $\sigma$) when $np \ge 5$ and $n(1-p) \ge 5$.

## 7 Statistical Inference for the Population Mean $\mu$

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. The *observed* sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ is an estimate of the mean of the population $\mu$.
- $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$ is the standard deviation of the sample. The *observed* standard deviation of the sample $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}$ is an estimate of the standard deviation of the population $\sigma$.
- The standard deviation of the sample mean is $\text{Std Dev}(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$, where $\sigma$ is the standard deviation of the population.
- If $n$ is large (say, 30 or more), then $\dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ is approximately a standard Normal RV.
- If $n$ is small (say, less than 30) and the population distribution is "well-behaved", then $\dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ obeys a $t$-distribution with $n-1$ degrees of freedom (dof).
- For $n \ge 30$ an $\alpha$% confidence interval for the real mean $\mu$ is $\left[\bar{x} - c \times \dfrac{s}{\sqrt{n}}, \bar{x} + c \times \dfrac{s}{\sqrt{n}}\right]$, where $c$ can be found by solving $P(-c \le Z \le c) = \alpha/100$ with $Z$ being a standard Normal RV. For example:

  For $\alpha = .90$, $c = 1.645$; for $\alpha = .95$, $c = 1.960$; for $\alpha = .98$, $c = 2.326$; for $\alpha = .99$, $c = 2.576$.

- For $n < 30$ and a "well-behaved" population distribution, an $\alpha\%$ confidence interval for the real mean $\mu$ is $\left[\bar{x} - c \times \frac{s}{\sqrt{n}}, \bar{x} + c \times \frac{s}{\sqrt{n}}\right]$, where $c$ satisfies that $P(-c \leq T \leq c) = \alpha/100$ with $T$ a RV that has a $t$-distribution with $n-1$ dof.

- To construct an $\alpha\%$ confidence interval that is within (plus or minus) $L$ of the actual mean, the required sample size is $n = \frac{c^2 s^2}{L^2}$, where $c$ satisfies that $P(-c \leq Z \leq c) = \alpha/100$ if $Z$ is a standard Normal RV.

# 8 Regression

- $n =$ number of data points

- $k =$ number of explanatory (independent) variables

- Based on observed data
  $(y_1, x_{11}, \ldots, x_{k1})$
  $\vdots$
  $(y_n, x_{1n}, \ldots, x_{kn})$

- Population relation: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \epsilon_i$ where $\epsilon_i$ is $N(0, \sigma)$

- $\hat{y}_i = b_0 + b_1 x_{1i} + \ldots + b_k x_{ki}$ is the predicted value

- $b_j$ is the regression coefficient and an estimate of $\beta_j$, $j = 0, 1, \ldots, k$

- $s_{b_j}$ is the standard deviation of $b_j$

- $e_i = y_i - \hat{y}_i$ is the residual

- An $\alpha\%$ confidence interval for $\beta_j$ is $[b_j - c \times s_{b_j}, b_j + c \times s_{b_j}]$
  where $c$ satisfies that $P(-c \leq T \leq c) = \alpha/100$ if $T$ obeys a $t$-distribution with dof $= n - k - 1$

- The $t$-statistic is $t_{b_j} = \frac{b_j}{s_{b_j}}$

- Checklist for evaluating a linear regression model: (i) linearity, (ii) signs of regression coefficients, (iii) significance of independent variables, (iv) $R^2$, (v) normality of the residuals, (vi) heteroscedasticity, (vii) autocorrelation, and (viii) multicolinearity.