Five number summary:

min, Q_1, median, Q_3, max

Q_1: median of smallest half

Q_3: median of largest half

Fourth spread

$f_s = Q_3 - Q_1$

Outliers

x_i is an outlier if its distance from the closest fourth (Q_1 or Q_3) is $> 1.5 f_s$

Sample variance

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{n-1} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)$$

Sample standard deviation

$s = \sqrt{s^2}$

Permutations

$$P_{k,n} = \frac{n!}{(n-k)!}$$

Combinations

$$C_{k,n} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Addition rule

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

Multiplication rule

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

Independent events

A and B are independent if $P(B|A) = P(B)$, equivalently $P(A \cap B) = P(A)P(B)$

Law of Total Probability

$A_1, \dots, A_k$ mutually exclusive & exhaustive:

$P(B) = P(A_1 \cap B) + \cdots + P(A_k \cap B)$

Special case: $P(E) + P(E') = 1$

De Morgan's laws

$(A \cup B)' = A' \cap B'$

$(A \cap B)' = A' \cup B'$

Expected value for a discrete r.v.

$E(X) = \mu_X = \sum x p(x)$

$E(h(X)) = \sum h(x) p(x)$

Expected value for a continuous r.v.

$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx$

$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx$

Variance and standard deviation

$V(X) = \sigma_X^2 = E(X^2) - E(X)^2$

$\sigma_X = \sqrt{V(X)}$

Rule for expected value

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

Rule for variance

$V(aX + b) = a^2 V(X)$

Binomial distribution

$X \sim \text{Bin}(n, p)$:

$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \dots, n$

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

Hypergeometric distribution

$n =$ sample size, $N =$ population size,

$M =$ number of successes in population

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$E(X) = n \cdot \frac{M}{N}$, $V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$

Poisson distribution

$X \sim \text{Poisson}(\lambda)$:

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1 \dots$

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

Percentiles

$\eta = 100p$ th percentile of X (continuous r.v.):

$P(X \leq \eta) = p$

Normal distribution

If $X \sim N(\mu, \sigma)$ then $\frac{X - \mu}{\sigma} \sim N(0,1)$

For $Z \sim N(0,1)$ set $\Phi(z) = P(Z \leq z)$

$\Phi(z_\alpha) = 1 - \alpha$

Statistics

$X_1, \dots, X_n$ random sample:

$\overline{X} = \frac{1}{n} \sum X_i$ (sample mean)

$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$ (sample variance)

Sampling distributions

$X_1, \dots, X_n$ random sample,

$X_i \sim$ distribution with mean μ and std. dev. σ:

$E(\overline{X}) = \mu$, $V(\overline{X}) = \sigma^2/n$

CLT: $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad (n > 30)$

Regression and Correlation

$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$

$S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$

$S_{xy} = \Sigma(x_i y_i) - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$

$SSE = \Sigma y_i^2 - \hat{\beta}_0 \Sigma y_i - \hat{\beta}_1 \Sigma x_i y_i$

$SST = S_{yy}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n}$

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, $\quad r^2 = 1 - \frac{SSE}{SST}$

$s^2 = \frac{SSE}{n-2}$

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

(α = significance level) (100$(1-\alpha)$% confidence level)

One mean

$H_0 : \mu = \mu_0$

$$z^* = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1) \qquad \text{(if } n > 30\text{)}$$

$$t^* = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \qquad \text{(if data normally distr.)}$$

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \qquad \text{(if } n > 30\text{)}$$

$$\overline{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \qquad \text{(if data normally distr.)}$$

Difference of two means

$H_0 : \mu_1 - \mu_2 = \Delta_0$

$$z^* = \frac{\overline{x}_1 - \overline{x}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \qquad \text{(if } n_1 > 30 \text{ and } n_2 > 30\text{)}$$

$$t^* = \frac{\overline{x}_1 - \overline{x}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu \qquad \text{(if data normally distr.)}$$

$$\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{(if } n_1 > 30 \text{ and } n_2 > 30\text{)}$$

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{(if data normally distr.)}$$

$$\text{where } \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

One proportion

$H_0 : p = p_0$

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1) \qquad \text{(if } n \text{ large)}$$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Difference of two proportions

$H_0 : p_1 - p_2 = \Delta_0$

$$z^* = \frac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1) \qquad \text{(if } n_1,\ n_2 \text{ large)}$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

One variance

$H_0 : \sigma^2 = \sigma_0^2$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1} \qquad \text{(if data normally distr.)}$$

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}},\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

Ratio of two variances

$H_0 : \sigma_1^2 = \sigma_2^2$

$$f^* = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} \qquad \text{(if data normally distr.)}$$

property of critical F-values: $F_{1-\alpha/2,\nu_1,\nu_2} = \dfrac{1}{F_{\alpha/2,\nu_2,\nu_1}}$

Slope of regression line

$H_0 : \beta_1 = \beta_{10}$

$$t^* = \frac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{S_{xx}}} \sim t_{n-2} \qquad \text{(if data normally distr.)}$$

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{s}{\sqrt{S_{xx}}}$$

Correlation coefficient

$H_0 : \rho = 0$

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \qquad \text{(if data normally distr.)}$$

Test of normality (Ryan-Joiner)

H_0 : population distribution is normal

test statistic: correlation coefficient r from probability plot

if $r < r_c$, reject H_0