Statistical Formulae

1. Calculation formulae

For a sample $x_1, x_2, \ldots, x_n$

- Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

2. The Binomial distribution

Assume $\boxed{X \sim \text{Bin}(n, \pi)}$

- **domain of variation :** $S_X = \{0, 1, \ldots, n\}$
- **probability mass function (pmf) :**

$$p(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \qquad \text{for } x \in S_X$$

Note that $\binom{n}{x} = {}^nC_x = \frac{n!}{x!(n-x)!}$.

- **cumulative distribution function (cdf) :**

$$F(x) = \sum_{k=0}^{\lfloor x \rfloor}\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

(where $\lfloor x \rfloor$ denotes the integer part of x).

- **expectation :**

$$\mathbb{E}(X) = n\pi$$

- **variance :**

$$\mathbb{V}\text{ar}(X) = n\pi(1-\pi)$$

3. The Poisson distribution

Assume $\boxed{X \sim \mathcal{P}(\lambda)}$

- **domain of variation :** $S_X = \{0, 1, 2, \ldots\}$
- **probability mass function (pmf) :**

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad \text{for } x \in S_X$$

- **cumulative distribution function (cdf) :**

$$F(x) = e^{-\lambda}\sum_{k=0}^{\lfloor x \rfloor}\frac{\lambda^k}{k!}$$

(where $\lfloor x \rfloor$ denotes the integer part of x). See also the attached Poisson table.

- **expectation :**

$$\mathbb{E}(X) = \lambda$$

- **variance :**

$$\mathbb{V}\text{ar}(X) = \lambda$$

4. The Uniform distribution

Assume $\boxed{X \sim U_{[\alpha,\beta]}}$

- **domain of variation :** $S_X = [\alpha, \beta]$
- **probability density function (pdf) :**

$$f(x) = \frac{1}{\beta - \alpha}, \qquad \text{for } x \in S_X$$

- **cumulative distribution function (cdf) :**

$$F(x) = \frac{x - \alpha}{\beta - \alpha}, \qquad \text{for } x \in S_X$$

- **expectation :**

$$\mathbb{E}(X) = \frac{\alpha + \beta}{2}$$

- **variance :**

$$\mathbb{V}\text{ar}(X) = \frac{(\beta - \alpha)^2}{12}$$

5. The Exponential distribution

Assume $\boxed{X \sim \text{Exp}(\mu)}$

- **domain of variation :** $S_X = [0, +\infty)$
- **probability density function (pdf) :**

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \qquad \text{for } x \in S_X$$

- **cumulative distribution function (cdf) :**

$$F(x) = 1 - e^{-\frac{x}{\mu}}, \qquad \text{for } x \in S_X$$

- **expectation :**

$$\mathbb{E}(X) = \mu$$

- **variance :**

$$\mathbb{V}\text{ar}(X) = \mu^2$$

6. The Normal distribution

Assume $\boxed{X \sim \mathcal{N}(\mu, \sigma)}$

- **domain of variation :** $S_X = (-\infty, +\infty)$
- **probability density function (pdf) :**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \qquad \text{for } x \in S_X$$

- **cumulative distribution function (cdf) :**

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}\, dy, \qquad \text{for } x \in S_X$$

  (no closed form)
- **expectation :**

$$\mathbb{E}(X) = \mu$$

- **variance :**

$$\mathbb{V}\text{ar}(X) = \sigma^2$$

7. Sampling distributions

7.1. **Sample mean.**

7.1.1. *known variance.* Let $\bar{X}$ be the sample average from a random sample of size n from a population with mean μ and standard deviation σ. Under appropriate conditions,

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

(exact result if the population distribution is normal, approximate result if the population distribution is not normal but $n > 30$)

7.1.2. *unknown variance.* Let $\bar{X}$ and S be the sample average and standard deviation from a random sample of size n from a normal population with mean μ. Under appropriate conditions,

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}$$

If the population is not normal but n is large enough ($n > 40$), we can also write

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim \mathcal{N}(0,1)$$

approximately

7.2. **Sample proportion.**

Let $\hat{p}$ be the sample proportion of 'successes' where the number of trials is n and the true probability of a success is π. Under appropriate conditions,

$$\sqrt{n}\,\frac{\hat{p}-\pi}{\sqrt{\pi(1-\pi)}} \sim N(0,1)$$

approximately when $n\pi(1-\pi) > 5$

7.3. **Sample variance.**

Let S^2 be the sample variance from a random sample of size n from a normal population with variance σ. Under appropriate conditions,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

7.4. **Difference in sample means.**

7.4.1. *variances σ_1^2 and σ_2^2 known.* For two independent samples of size n_1 and n_2 from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, let $\bar{X}_i$ be the sample average of sample i for $i = 1$ and 2. Under appropriate conditions,

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

(exact result if both population distributions are normal, approximate result if they are not but $n_1, n_2 > 30$)

7.4.2. *variances σ_1^2 and σ_2^2 unknown; $\sigma_1^2 = \sigma_2^2$.* For two independent samples of size n_1 and n_2 from two normal populations with means μ_1 and μ_2 respectively and common standard deviation σ, let $\bar{X}_i$ and S_i be the sample average and sample standard deviation of sample i for $i = 1$ and 2. Under appropriate conditions,

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where S_p is the pooled sample standard deviation,

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}.$$

7.4.3. *variances σ_1^2 and σ_2^2 unknown; $\sigma_1^2 \neq \sigma_2^2$.* For two independent samples of size n_1 and n_2 from two normal populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, let $\bar{X}_i$ and S_i be the sample average and sample standard deviation of sample i for $i = 1$ and 2. Under appropriate conditions,

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_\nu,$$

where

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

(rounded down to the nearest integer)

7.5. **Ratio of sample variances.** Let S_1^2 and S_2^2 be the sample variances from two independent random samples of size n_1 and n_2 from normal populations with variances σ_1^2 and σ_2^2 respectively. Under appropriate conditions,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathbf{F}_{n_1-1,n_2-1}$$

8. Simple linear regression

Consider the simple linear regression model $\quad Y = \beta_0 + \beta_1 X + \epsilon \quad$ where $\epsilon \sim \mathcal{N}(0, \sigma)$

The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where

$$S_{XY} = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \qquad S_{XX} = \sum_i (X_i - \bar{X})^2.$$

An estimator of σ is

$$S = \sqrt{\frac{\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2}}.$$

Under fixed design :

$$\sqrt{s_{xx}}\frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \sim t_{n-2}$$

Let x_0 denote the predictor value for a response yet to be observed :

i) a $100 \times (1-\alpha)\%$ confidence interval for the mean response at x_0 is

$$\left[\hat{y}(x_0) \pm s\, t_{n-2;1-\alpha/2}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}\right]$$

where $\hat{y}(x_0) = \hat{b}_0 + \hat{b}_1 x_0$;

ii) a $100 \times (1-\alpha)\%$ prediction interval for the response at x_0 is

$$\left[\hat{y}(x_0) \pm s\, t_{n-2;1-\alpha/2}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}\right]$$

9. ANOVA

- Total sum of squares :

$$SS_{\text{Tot}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2$$

- Treatment sum of squares :

$$SS_{\text{Tr}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$$

- Error sum of squares :

$$SS_{\text{Er}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

- Under the assumption of equality of means in each of the k groups of a one-way Analysis of Variance,

$$F = \frac{\text{MS}_{\text{Tr}}}{\text{MS}_{\text{Er}}} \sim \mathbf{F}_{k-1,n-k},$$

where n is the total number of observations.