Definitions

Two events A, B are **independent** iff $P(A \cap B) = P(A) \cdot P(B)$ and $P(A \mid B) = P(A)$. Two events are **mutually exclusive (disjoint)** iff $A \cap B = \emptyset$.

A **Bernoulli distribution** is used for an RV with a binary outcome and a probability of success p. A **geometric distribution** describes the number of Bernoulli trials X (each with probability of success p) until a success. A **binomial distribution** describes the number of successful events X in n Bernoulli trials. A **negative binomial distribution** describes the number of Bernoulli trials X until there are r successes. A **Poisson distribution** describes data rates (per unit time).

For a discrete RV, the **probability mass function (pmf/pdf)** is the function $f(x)$ such that $f(x) = P(X = x)$. The **cumuluative distribution function (CDF)** is the function $F(x)$ such that $F(x) = P(X \leq x)$. For a continuous RV, the **probability density function (PDF)** is the function $f(x)$ such that $P(a \leq X \leq b) = \int_a^b f(x)\,dx$.

A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ. It is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** $\hat{\theta}$. The average error of $\hat{\theta}$ is its **bias**. The **standard error** tells us about the quality of the estimator.

The **z-table** is typically used for confidence intervals and hypothesis tests on normally distributed data when $n \geq 30$ or when the population standard deviation σ is known. z_α is defined as $\Phi^{-1}(\alpha)$.

A **p-value** describes the probability that something as or more extreme than what we measured happens assuming H_0 is true.

A **paired sample** is a set of data where each observation in one group is directly linked to another. We're interested in the differences between the groups, $d = x_1 - x_2$. n_d denotes the number of differences. The average difference is $\bar{d} = \bar{x}_1 - \bar{x}_2$. For **differences between proportions**, the proportion for sample n is represented as $\hat{p}_n$, and its variance is $\hat{p}_n(1 - \hat{p}_n)$.

The **coefficient of determination** R^2 is a measure in $[0, 1]$ describing how much of the variance of y is explainable by x. For simple linear regression, $R^2 = r^2$. In this case, to determine the correlation r using R^2, use $r = \text{sign}(\hat{\beta}_1)\sqrt{R^2}$.

	H_0 is actually true	H_0 is actually false
Reject H_0	Type I error (false positive) (probability: α)	True positive ($1 - \beta$ = power)
Fail to reject H_0	True negative (probability: $1 - \alpha$)	Type II error (false negative) (probability: β)

Formulas

Sample mean	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	Union probability	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Definition of S_{xy}	$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$	Permutations ordered, k out of n objects	${}_nP_k = \frac{n!}{(n-k)!}$
Sample variance	$s^2 = \frac{S_{xx}}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$	Combinations unordered, k out of n objects	${}_nC_k = \frac{n!}{k!(n-k)!}$
Standard deviation	$s = \sqrt{s^2}$	Multiplication rule	$P(A \cap B) = P(A \mid B)P(B)$
Position of pth Percentile	$\begin{cases} (p\%)n + \frac{1}{2} & (p\%)n \in \mathbb{Z} \\ \lceil (p\%)n \rceil & \text{otherwise} \end{cases}$	Conditional probability	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
Interquartile Range	$\text{IQR} = Q_3 - Q_1$	Outlier cutoff	$(Q_1 - 1.5\,\text{IQR}, Q_3 + 1.5\,\text{IQR})$
Correlation r	$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}}$	Law of Total Probability	$P(B) = \sum_{i=1}^k P(B \mid A_i)P(A_i)$

Bayes Theorem	$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^k P(B \mid A_i)P(A_i)}$

Discrete probabilities with CDF	$P(a \leq X \leq b)$	$=$	$F(b) - F(a-1)$
	$P(a < X \leq b)$	$=$	$F(b) - F(a)$
	$P(a \leq X < b)$	$=$	$F(b-1) - F(a-1)$
	$P(a < x < b)$	$=$	$F(b-1) - F(a)$
	$P(x = a)$	$=$	$F(a) - F(a-1)$
Expected value (μ)	$E(X)$	$=$	$\sum_{i=1}^n x_i f(x_i)$
	$E(h(X))$	$=$	$\sum_{i=1}^n h(x_i) f(x_i)$
	$E(aX + b)$	$=$	$a \cdot E(X) + b$
Variance	$V(X)$	$=$	$\sum_n (x - \mu)^2 f(x)$
		$=$	$E((X - \mu)^2)$
		$=$	$E(X^2) - (E(X))^2$
	$V(h(X))$	$=$	$E(h(X)^2) - (E(h(X)))^2$
	$V(aX + b)$	$=$	$a^2 V(X)$

Gamma function	$\Gamma(x) = (x-1)! = \int_0^\infty t^{x-1}e^{-t}\,dt$
Continuous RV expected value	$\mu_x = E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx$
Continuous RV variance	$\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx = E(X^2) - (E(X))^2$
nth percentile of an RV	$P(X \le v) = \displaystyle\int_{-\infty}^{v} f(x)\,dx \overset{\text{set}}{=} n\%$
Central Limit Theorem for $X_1, X_2, \ldots, X_n$	$\bar{X} = N\left(\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\right)$
Bias	$\mathrm{Bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$
Mean squared error	$\mathrm{MSE} = E((\hat{\theta} - \theta)^2) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})$
Standard error	$\mathrm{StdError}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})} = \dfrac{\sigma}{\sqrt{n}}$
$1-\alpha$ confidence interval for μ (z-test)	$\bar{x} \pm z_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$
$1-\alpha$ confidence interval for μ (t-test)	$\bar{x} \pm t_{\alpha/2,\,n-1}\left(\dfrac{s}{\sqrt{n}}\right)$
95% confidence interval for μ	$\left(\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$
Normal dist. standardization (z test statistic)	$z = \dfrac{x-\mu}{\sigma} = \dfrac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}}$
t test statistic	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Left-tailed test critical value	$z_\alpha = \Phi^{-1}(\alpha)$ or $-t_{\alpha,\,n-1}$
Right-tailed test critical value	$-z_\alpha$ or $z_{1-\alpha}$ or $t_{\alpha,\,n-1}$
Two-tailed test critical value	$z_{\alpha/2}$ or $\pm t_{\alpha/2,\,n-1}$
H_0 rejection	$\lvert\text{Test stat}\rvert > \lvert\text{Critical val}\rvert$ or p-value $\le \alpha$
Two-sample degrees of freedom	$v = \min(n_1 - 1,\ n_2 - 1)$
$1-\alpha$ confidence interval for $\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\,v}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
t_{stat} when H_0: $\mu_1 - \mu_2 = \Delta_0$	$t_{\text{stat}} = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$
$1-\alpha$ confidence interval for μ_d	$\bar{d} \pm t_{\alpha/2,\,n_d-1}\left(\dfrac{s_d}{\sqrt{n_d}}\right)$
t_{stat} when H_0: $\mu_d = \Delta_0$	$t_{\text{stat}} = \dfrac{\bar{d} - \Delta_0}{s_d/\sqrt{n_d}}$
$1-\alpha$ confidence interval for $p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
z_{stat} when H_0: $p_1 - p_2 = \Delta_0$	$z_{\text{stat}} = \dfrac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ where $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
Least squares regression line	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
True slope $\hat{\beta}_1$	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$
True intercept $\hat{\beta}_0$	$\hat{\beta}_0 = \dfrac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x^i}{n} = \bar{y} - \hat{\beta}_1\bar{x}$
Least squares estimate of variance	$\sigma^2 = \dfrac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \dfrac{\mathrm{SSE}}{n-2} = \dfrac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$
Coefficient of determination R^2	$R^2 = 1 - \dfrac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \dfrac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
Correlation r for simple linear regression	$r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \mathrm{sign}(\hat{\beta}_1) \cdot \sqrt{R^2}$