Applied Bioostatistics I - formulas

Axioms of Kolmogorov:

$A(\textit{probability measure}) \subset \Omega(\textit{sample space}) : P[A] \in [0,1]$:

- $0 \leq P[A] \leq 1$ for every event $A \subset \Omega$
- $P[\Omega] = 1$
- $P[A \cup B] = P[A] + P[B]$ for *disjoint* event A and B.

for A, B events

De Morgan's laws: $(A \cap B)^c = A^c \cup B^c$ **and** $(A \cup B)^c = A^c \cap B^c$

Probability of unions: $P[A \cup B] = P[A] + P[B] - P[A \cap B]$

General: $A_1, A_2, \ldots, A_n$ events:
$P\left[A_1 \cup A_2 \cup \ldots \cup A_n\right] = \sum_{i_1=1}^{n} P\left[A_{i_1}\right] - \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^{n} P\left[A_{i_1} \cap A_{i_2}\right] + \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^{n} P\left[A_{i_1} \cap A_{i_2} \cap A_{i_3}\right] - \ldots$

Discrete probability spaces

$A \subset \Omega : P[A] = \sum_{i\omega_i \in A} P\left[\{\omega_i\}\right]$
$P[\Omega] = \sum_{i \geq 1} P\left[\{\omega_i\}\right]$
Ω is finite often: $P\left[\{\omega_i\}\right] = 1/|\Omega|$

Independence if: $P[A \cap B] = P[A] \cdot P[B]$

Conditional probability: $P[A|B] = \frac{P[A \cap B]}{P[B]}$ **, given** $P[B] > 0$

Law of total probability

for disjoint $B_1, B_2, \ldots, B_k$ and $B_1, B_2, \ldots, B_k = \Omega$:
$P[A] = \sum_{i=1}^{k} P\left[A \cap B_i\right] = \sum_{i=1}^{k} P[A|B_i] P\left[B_i\right]$

Bayes' theorem:

$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]}$, given $P[A] > 0$ and $P[B] > 0$
Total probability: $P\left[B_i|A\right] = \frac{P[A \cap B_i]}{P[A]} = \frac{P[A|B_i]P[B_i]}{\sum_{j=1}^{k} P[A|B_j]P[B_j]}$

Cumulative distribution function (CDF) of a random variable X: $F_X(x) := P[X \leq x]$

continuous: $F(x) = \int_{-\infty}^{x} f(u) \mathrm{d}u$

Discrete random variables

$X : \Omega \to \{x_1, x_2, \ldots\}$; **probability mass function** $p\left(x_k\right) := P\left[X = x_k\right]$; $A \subset \{x_1, x_2, \ldots\}$:

- $P[X \in A] = \sum_{k: x_\in \in A} p\left(x_k\right)$
- $\sum_k p\left(x_k\right) = 1$
- CDF : $F_X(x) = P[X \leq x] = \sum_{k: x_k \leq x} p\left(x_k\right)$

Expectation value: $E[X] := \sum_k x_k p\left(x_k\right)$

continuous: $E[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$

Variance: $\mathrm{Var}(X) := \sum_k \left(x_k - E[X]\right)^2 p\left(x_k\right) = \mathrm{SD}^2(X)$ **: Standard Deviation** ^{2}

continuous: $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$

Bernoulli distribution $X \in \{0,1\}$ $X \sim$ **Bernoulli(** π) **,** $\pi := P[X=1]$

BINOMial distribution $X \in \{0,1,\ldots,n\}$ $X \sim \text{Bin}(n,\pi), n \in \mathbb{N}, \pi \in (0,1)$

- $p(x) = P[X=x] = \left(\begin{array}{c} n \\ x \end{array}\right) \pi^x (1-\pi)^{n-x}$
- $E[X] = n\pi, \text{Var}(X) = n\pi(1-\pi)$
- Approx.: $(n\pi > 5$ & $n(1-\pi) > 5)$? $X \approx \mathcal{N}(n\pi, n\pi(1-\pi))$ (only CDF) ; $(n \geq 100$ & $\pi \leq 0.01$ & $n\pi \leq 20)$? $X \approx \text{Pois}(n\pi)$

POISson distribution $X \in \mathbb{N}$ $X \sim \text{Pois}(\lambda), \lambda > 0$

- $p(x) = P[X=x] = \frac{e^{-\lambda}\lambda^x}{x!}$
- $E[X] = \lambda, \text{Var}(X) = \lambda$
- CDF: $F(x;\lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$
- $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$ independent: $X+Y \sim \text{Pois}(\lambda_1+\lambda_2)$

UNIForm distribution $X \sim \mathcal{U}([a,b])$

- $f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$
- $E[X] = \frac{b+a}{2}, \text{Var}(X) = \frac{(b-a)^2}{12}$

NORMal distribution $X \sim \mathcal{N}\left(\mu, \sigma^2\right), \mu \in \mathbb{R}, \sigma^2 > 0$

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, x \in \mathbb{R}$

Standard normal distribution $Z \sim \mathcal{N}(0,1)$

- $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \int_{-\infty}^{z} \varphi(t)dt$
- $Z = aX + b \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right)$
- $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

EXPonential distribution $X \sim \text{Exp}(\lambda), \lambda > 0$

- $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$
- $E[X] = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}$
- Haldane's model: (position of the first crossover) $X \sim \text{Exp}(1)$

discrete	continuous
$E[X] = \sum_{k\geq 1} x_k p(x_k)$	$E[X] = \int_{-\infty}^{\infty} x f(x)dx$
$\text{Var}(X) = \sum_{k\geq 1}(x_k - E[X])^2 p(x_k) = E\left[X^2\right] - (E[X])^2$	$\text{Var}(X) = \int_{-\infty}^{\infty}(x - E[X])^2 f(x)dx$

R function naming

- p or "probability", the cumulative distribution function (c. d. f.)
- q for "quantile", the inverse c. d. f.
- d for "density", the density function (p. f. or p. d. f.)
- r for "random", a random variable having the specified distribution

Discrete multivariate distributions

Let $X : \Omega \to W_x$ and $Y : \Omega \to W_Y$ be discrete random variables
Joint Cumulative Distribution Function: $F_{X,Y}(x,y) := P[X \leq x, Y \leq y]$
Joint Probability Mass Function: $p_{X,Y}(x,y) := P[X=x, Y=y], x \in W_X, y \in W_Y$
Marginal Probability Mass Function: $p_X(x) = P[X=x] = \sum_{y \in W_Y} p_{X,Y}(x,y)$
Independence IF: $p_{X,Y}(x,y) = p_X(x)p_Y(y)$
Conditional Probability Mass function: $p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

Continuous multivariate distributions

Let $X \to \mathbb{R}$ and $Y \to \mathbb{R}$ be continuous random variables
Joint cumulative distribution function: $F_{X,Y}(x,y) := P[X \leq x, Y \leq y]$
Joint probability density: $f_{X,Y}(x,y) := \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{X,Y}(x,y)$
$P[a \leq X \leq b, c \leq Y \leq d] = \int_a^b \int_c^d f_{X,Y}(x,y)dydx \ (a < b, c < d)$
Marginal probability density: $f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y$
Independence IF: $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
Conditional probability density: $f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

Covariance	Correlation
$\mathrm{Cov}(X,Y) := E[(X - E[X])(Y - E[Y])]$	$\rho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$

- if X, Y independent => $\mathrm{Cov}(X,Y) = 0$, $\rho_{XY} = 0$, $E[XY] = E[X] \cdot E[Y]$ and $\mathrm{Cov}(X,Y) = 0$ (the other direction is not true!)
- $-1 \leq \rho_{XY} \leq 1$
- $\rho_{XY} = 1$ if $Y = a + bX$ for some $b > 0$
- $\rho xy = -1$ if $Y = a + bX$ for some $b < 0$
- $E[X + Y] = E[X] + E[Y]$
- $E[aX] = aE[X]$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$
- $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$

Descriptive Statistics: Model: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F_X(\cdot)$ **, Sample:** $x_1, x_2, \ldots, x_n$

Sample MEAN: $\overline{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^n x_i = S_n/n \to \mu = E[X]$ if $n \to \infty$ (consistent/unbiased estimator for the true mean) Sample VARiance: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ (s_x : sample Standard Deviation) $s_x^2 \to \sigma^2 = \mathrm{Var}(X)$ if $n \to \infty$
$E\left[s_x^2\right] = \sigma^2$ (consistent/unbiased estimator for the true variance)
MEDIAN: $(x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$: $m = \begin{cases} x_((n+1)/2), \ n \ is \ odd, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), \ otherwise \end{cases}$
Empirical α QUANTILE: $q_\alpha = x_{(\alpha(n-1)+1)}$ if $\alpha \cdot (n-1)$ is an integer; otherwise $(x_{(\lfloor\alpha(n-1)\rfloor+1)} + x_{(\lceil\alpha(n-1)\rceil+1)})/2$
random variable X : value m such that $P[X \leq m] \geq \alpha$ and $P[X \geq m] \geq 1 - \alpha$

Histogram: bins $h_k := \#\{i|x_i \in (c_{k-1}, c_k]\}$ **density** $\frac{h_k}{n(c_k - c_{k-1})}$ **, Box plot IQR ("interquartile range"):** $q_{0.75} - q_{0.25}$

Kernel density estimation

Given a set of points $x_1, x_2, \ldots, x_n$, the kernel density estimator for the generating distribution is
$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$ (kernel function:arbitrary positive symmetric, h: bandwidth)
* Uniform/rectangular kernel: $K \sim \mathcal{U}\left(\left[-\frac{1}{2}, \frac{1}{2}\right]\right)$ (same weight for all points)
* Gaussian kernel: $K \sim \mathcal{N}(0,1)$ (less weight to far apart points)

Empirical cumulative distribution function (ECDF): $\hat{F}(x) = \frac{\#\{k|x_k \leq x\}}{n}$
Empirical CORrelation: $r = \frac{s_{xy}}{s_x s_y} \in [-1,1]$, empirical COVariance: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$
Linear dependence between 2 samples $\{x_i\}$ and $\{y_i\}$ * $r = +1$ if $y_i = a + bx_i$ for some $b > 0$ * $r = -1$ if $y_i = a + bx_i$ for some $b < 0$

Standard error of the mean (SEM) = Standard Deviation of the mean

Natural estimator for $\sigma\left(\overline{X}_n\right)$: $\mathrm{se}_{\overline{x}} = \frac{s_x}{\sqrt{n}}$; s_x is the empirical standard deviation

Law of large numbers: X **random variable and** $E[X] = \mu$**, and** $X_1, X_2, \ldots, X_n$ **i.i.d. from** X. **=>** $\overline{X}_n \to \mu$ **as** $n \to \infty$

ECDF $\to$ CDF as $n \to \infty$

Central Limit Theorem

Let X be random variable with expectation value μ and variance σ^2, and $X_1, X_2, \ldots, X_n$ i.i.d. from X .
Then $\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ for large n. : $E\left[\overline{X}_n\right] = \mu, \sigma\left(\overline{X}_n\right) = \frac{\sigma}{\sqrt{n}} \to 0$ as $n \to \infty$
$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \approx \mathcal{N}(0,1)$ for large n
$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and $Y = a + bX$: $Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right)$

Confidence interval with confidence level $1-\alpha, \frac{1}{2}<\alpha<1$

$\left[\overline{X}_n-\Phi^{-1}(1-\alpha/2)\cdot\frac{s_x}{\sqrt{n}}, \overline{X}_n+\Phi^{-1}(1-\alpha/2)\cdot\frac{s_x}{\sqrt{n}}\right]$

Maximum likelihood estimation (MLE) for discrete distributions with measurements $X_1, X_2, \ldots, X_n$: i.i.ds

probability mass function $p(x;\theta)$: parameterized by θ
Likelihood $L(\theta):=\prod_{i=1}^n p\left(x_i;\theta\right)$
Log-likelihood $\ell(\theta):=\log(L(\theta))=\sum_{i=1}^n \log\left(p\left(x_i;\theta\right)\right)$
Maxiumum likelihood estimator (MLE) for θ : $\hat{\theta}=$ value of θ for which ℓ attains its maximum; suitable for large samples

MLE for continuous distributions with probability density $f(x;\theta)$: parameterized by θ

$L(\theta):=\prod_{i=1}^n f\left(x_i;\theta\right)$
$\ell(\theta):=\log(L(\theta))=\sum_{i=1}^n \log\left(f\left(x_i;\theta\right)\right)$

MLE for Poisson distribution

$L(\lambda)=\prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$
$\ell(\lambda)=\sum_{i=1}^n\left[x_i\log(\lambda)-\lambda-\log\left(x_i!\right)\right]$
$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^n x_i=\overline{x}$
confidence intervals: $\left[\hat{\lambda}-\Phi^{-1}(0.975)\frac{s_x}{\sqrt{n}}, \hat{\lambda}+\Phi^{-1}(0.975)\frac{s_x}{\sqrt{n}}\right]$

MLE for Normal distribution $\mathcal{N}\left(\mu,\sigma^2\right)$

$\hat{\mu}=\overline{X}, \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n\left(x_i-\overline{x}\right)^2$

MLE for Exponential distribution $E\times p(\lambda)$

$\hat{\lambda}=\frac{1}{\overline{x}}$
confidence interval: $\left[\hat{\lambda}\left(1-\frac{\Phi^{-1}(0.975)}{\sqrt{n}}\right), \hat{\lambda}\left(1+\frac{\Phi^{-1}(0.975)}{\sqrt{n}}\right)\right]$

Bayesian estimation approach: parameter θ as random

Likelihood as conditional probability: $L(\theta)=p_{X|\Theta=\theta}(x)=P[X=x|\Theta=\theta]$
$P[\Theta=\theta|X=x]=\frac{P[X=x|\Theta=\theta]\cdot P[\Theta=\theta]}{P[X=x]}$: *posterior* $=\frac{\textit{likelihood}\cdot\textit{prior}}{\textit{evidence}}$
Maximum a posteriori (MAP) estimator: $\hat{\theta}$ that maximizes the *posterior* $P[\Theta=\theta|X=x]$, suitable for small samples

Bayesian estimation of continuous parameter with density $f_{\ominus}(\theta)$

$f_{\Theta|X=x}(\theta)=\frac{f_{X|\Theta=\theta}(x)\cdot f_{\ominus}(\theta)}{f_X(x)}$

In large sample limit, $n\to\infty$: MAP (Bayesian) estimate converges to ML (frequentist) estimate

Statistical Hypothesis Testing

1) Model: choose distribution describing your data. Formulate claim you want to prove.

2) Null hypothesis: choose the H_0(*null hypothesis*) , H_A(*alternative hypothesis*) and their distribution parameters

3) Test statistic: based on your sample data

4) Choose significance level: e.g. $\alpha=5\%$

5) Range of rejection K such that $P[X\in K]\leq\alpha$ under H_0
   reject H_0 if $X\in K$
6) Test decision: reject H_0 if $X\in K$ otherwise fail to reject.

		Decision	

		Decision	
		H_0	H_A
Truth	H_0	true negative (TN)	type I error (FP)
	H_A	type II error (FN)	true positive (TP)

- Significance level $\alpha : \alpha = P[FP|H_0]$
- Power $1-\beta : \beta = P[FN|H_A]$; higher Significance(smaller α) => lower Power ; bigger α or larger sample => higher Power

P-value

(Def.) The *p-value* is the smallest significance level α for which we reject a null hypothesis for the given data set.
(Alt.) The *p-value* is the probability under the null hypothesis to find the actual outcome or a more extreme one.

Test using the normal approximation $X \approx \mathcal{N}\left(n\pi_0, n\pi_0\left(1-\pi_0\right)\right)$

Test statistic: $Z = \frac{X-n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$ Distribution of Z under $H_0 : Z \approx \mathcal{N}(0,1)$

Paired-samples (or one-sample) t test with model $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$

$H_0 : \mu = \mu_0 = 0$; $H_{\text{A}} : \mu \neq \mu_0$
Test statistic: $T = \frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{s_x}$ with T distribution | H_0 with (n - 1) degrees of freedom

Student's T distribution $T \sim t_m$ **with** m **"degrees of freedom", symetry** $t_{m,\alpha} = -t_{m,1-\alpha}$

Range of rejection: $K = \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right)$

Confidence Interval for μ **with confidence level** $1-\alpha$

$I = \{\mu_0 |$ null hypothesis $H_0 : \mu = \mu_0$ is not rejected $\}$
$H_{\text{A}} : \mu \neq \mu_0 \Rightarrow I = \left[\overline{x} - t_{n-1,1-\alpha/2}\frac{s_x}{\sqrt{n}}, \overline{x} + t_{n-1,1-\alpha/2}\frac{s_x}{\sqrt{n}}\right]$
$H_{\text{A}} : \mu < \mu_0 \Rightarrow I = \left(-\infty, \overline{x} + t_{n-1,1-\alpha}\frac{s_x}{\sqrt{n}}\right]$
$H_{\text{A}} : \mu > \mu_0 \Rightarrow I = \left[\overline{x} - t_{n-1,1-\alpha}\frac{s_x}{\sqrt{n}}, \infty\right)$

Sign Test: consider differences $X_i = Z_i - Y_i$ **i. i. d. with median** m

$H_0 : m = m_0 = 0$, $H_{\text{A}} : m \neq m_0$
Test statistic: $V = \#\{i | X_i > m_0\}$, V under $H_0 : V \sim \text{Bin}(n, 0.5)$
Range of rejection: $K = [0,c] \cup [n-c, n]$ such that $P_{H_0}[V \in K] \leq \alpha$ (significance level)
c determined by binomial distribution: $P_{H_0}[V \in K] = 2P_{H_0}[V \leq c]$

Wilcoxon Signed-Rank Test (wilcox.test): concider differences $X_i = Z_i - Y_i$ **i. i. d. with median** m

$H_0 : m = 0$, $H_{\text{A}} : m \neq 0$
Test statistic: $W = \sum_{i=1}^{n} \operatorname{sign}\left(X_i\right) R_i$, where R_i: rank of X_i order by absolute value $|X_i|$
Range of rejection: $K = (-\infty, 0.5-c] \cup [0.5+c, \infty)$ such that $P_{H_0}[W \in K] \leqslant \alpha$

T-Test

$X_1, \ldots, X_n \stackrel{\text{i.i.d}}{\cdot} \mathcal{N}\left(\mu_X, \sigma^2\right)$, $Y_1, \ldots, Y_m \stackrel{\text{i.i.d}}{\sim} \mathcal{N}\left(\mu_Y, \sigma^2\right)$, $H_0 : \mu_X = \mu_Y$, $H_{\text{A}} : \mu_X \neq \mu_Y$
Test statistic: $T = \frac{\overline{X}-\overline{Y}}{s_{\text{ pool}}\sqrt{1/n+1/m}}$, $s_{\text{pool}}^2 = \frac{1}{n+m-2}\left((n-1)s_x^2 + (m-1)s_y^2\right)$, T distribution under $H_0 : T \sim t_{n+m-2}$

Permutation Test: nonparametric test

$X_1, \ldots, X_n \overset{\text{i.d.d}}{\sim} F_X(\cdot)$, $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} F_Y(\cdot)$
$H_0 : F_X = F_Y$, $H_\text{A} : F_x \neq F_Y$
Test statistic: $D = \overline{X} - \overline{Y}$
Resampling: choose number of repetitions $N > 1000$
Randomly assign n values of $\{X_i\} \cup \{Y_i\}$ to "type I" and the rest m values to "type II"
Repeat N times
Range of rejection: $K = (-\infty, c_l] \cup [c_u, \infty)$, c_l : empirical $\alpha/2$ -quantile of resampling distribution, c_u : empirical $1 - \alpha/2$ -quantile of resampling distribution

Effect size Two samples: Experimental group $\{X_i\}_i$, control group $\{Y_i\}_i$, effect size $= \frac{\overline{X} - \overline{Y}}{s_{\text{pool}}}$

False Positive Rate: $FPR = E\left[\frac{FP}{FP+TN}\right] = E\left[\frac{V}{m_0}\right]$ **controlled by significance level** $\alpha = FPR$

		Decision		Total
		H_0	H_A	
Truth	H_0	True Negative U	type I error (FP): V	m_0
	H_A	type II error (FN)	True Tositive: S	$m - m_0$
Total		$m - R$	R	m

Family-Wise Error Rate: $\text{FWER} = P[1$ **or more type** $|$ **errors** $] = P[V \geq 1]$ **: n:20-50, errors are critical**

FWER controlled by experiment-wise type I error rate $\overline{\alpha}$
Test procedure that guarantees a FWER of (at most) $\overline{\alpha}$:
1. for each test case (e.g. gene), calculate p-value
2. adjust p-value
3. reject null hypotheses whose adjusted p-value is smaller than $\overline{\alpha}$; accept others

Controlling FWER

order p-values: $P_{(1)} \leq P_{(2)} \leq P_{(3)} \leq \ldots \leq P_{(m)}$

	Bonferroni method	Holm method
adjust p-values	$P_{\text{adj},i} = \min\{m \cdot P_i, 1\}$	$P_{\text{adj},(i)} = \max\left\{\min\left\{(m - i + 1) \cdot P_{(i)}, 1\right\}, P_{\text{adj},(i-1)}\right\}$

reject null hypotheses whose adjusted p-value is smaller than $\overline{\alpha}$; accept others : guaranteed FWER $\leq \overline{\alpha}$

Adjusted p-value

The adjusted p-value of a certain null hypothesis is the smallest experiment-wise type I error rate $\overline{\alpha}$ for which we reject this hypothesis for the given data set.

False discovery rate: $\text{FDR} = E\left[\frac{\text{FP}}{\text{FP+TP}}\right] = E\left[\frac{V}{R}\right]$ **: n > 500 ; looking for discovery**

1. for each test case, calculate p-value
2. adjust p-values to get corresponding q-values
3. reject null hypotheses whose q-value is smaller than $\overline{q}$; accept others

Controlling FDR: Benjamini-Hochberg method

order p-values: $P_{(1)} \leq P_{(2)} \leq P_{(3)} \leq \ldots \leq P_{(m)}$
adjust p-values to get q-values: $Q_{(i)} = \max\left\{\min\left\{\frac{m}{i} \cdot P_{(i)}, 1\right\}, Q_{(i-1)}\right\}$
reject null hypotheses whose q-value is smaller than $\overline{q}$; accept others
Procedure guarantees $FDR \leq \overline{q}$

Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + E_i$, $E_1, \ldots, E_n \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, $i = 1, \ldots, n$

Y_i : response variable ; x_i : explanatory variable ; E_i : error or noise variables
Residuals: $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
Residual Sum of Squares: $\text{RSS} = \sum_{i=1}^{n} R_i^2$
Minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ of RSS: unbiased estimators for the true β_0 and β_1
Estimate $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2$

Significance of explanatory variable: t-test

$H_0 : \beta_1 = 0$; $H_\text{A} : \beta_1 \neq 0$; $T = \frac{\hat{\beta}_1 - 0}{\widehat{\text{se}}\left(\hat{\beta}_1\right)}$; $H_0 : T \sim t_{n-2}$; $K = \left(-\infty, -t_{n-2,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2,1-\frac{\alpha}{2}}, \infty\right)$, $|T| > t_{n-2,1-\frac{\alpha}{2}}$: *reject*

Coefficient of determination $R^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y}\right)^2$

$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)} = \text{Cor}(Y,X) \frac{\text{Sd}(Y)}{\text{Sd}(X)}$, $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$; $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Fitting linear model to transformed data: $\log\left(Y_i\right) = \hat{\beta}_0 + \hat{\beta}_1 \log\left(X_i\right) + E_i$