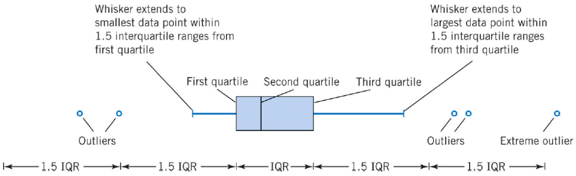
16. PCA

PCA diagonalize $p \times p$ corr coefficient matrix $r_{ij} = \sigma_{ij}/\sigma_i\sigma_j$.

18. Descriptive statistics

Box-and-Whisker plot:



Probability plot: $x_j (j - 0.5)/n/CDF(x_j)$

19. Sample mean and variance

sample **mean**: $E(\bar{X}) = \mu$

sample **variance**: $V(\bar{X}) = \sigma^2/n$

sample **std/standard error (SE)**: $\sigma/\sqrt{n}$

central limit theorem: the limiting form of the distribution of large n is the standard normal distribution: $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$

Two popluations: the sampling distribution of $\bar{X}_1 - \bar{X}_2$ with mean $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ and variance $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma_1^2/n_1 + \sigma_2^2/n_2$

Sampling distribution of a **difference** in sample means: $Z = ((\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

20. point estimator

Unbiased point estimator: $E(\hat{\Theta}) = \Theta$

Bias: $E(\hat{\Theta}) - \Theta$

Mean Squared Error: $MSE(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2 = V(\hat{\Theta}) + bias^2$

Methods of moment: kth moment of random variable is $E(X^k)$.

First moment: $\mu = \int x f(x) dx$; second moment: $\mu^2 + \sigma^2 = \int x^2 f(x) dx \to E(X^2) = Var(X) + E(X)^2$.

Moment estimators: $X_1, .., X_n$ with m unknown parameters. They are found by **equating first m population moments to first m sample moments**.

To estimate exponential distribution, 1st moment, $E(X) = \bar{x} = 1/\lambda$; higher moment, $E(X^p) = p!/\lambda^p$

Maximum likelihood: $L(\theta) = f(x_1, \theta).f(x_2, \theta)...f(x_n, \theta)$. Maximum likelihood estimator (MLE) of θ is the value of θ that maximize $L(\theta)$. Use logarithm: $I(\theta) = lnL(\theta)$

Exponential MLE: $dlnL(\lambda)/d\lambda = n/\lambda - \sum x_i = 0 \to \lambda = n/\sum x_i = 1/\bar{X}$

Bernoulli MLE: $\hat{p} = \sum x_i/n$

Normal MLE for μ: $dlnL(\mu)/d\mu = \sum(x_i - \mu)/\sigma^2 = 0 \to \hat{\mu} = \bar{X}$

MLE for poisson distribution: $dlnf(x_1, .., x_n|\lambda)/d\lambda = -n + \sum x_i/\lambda = 0 \to \lambda = \sum x_i/n$

Sample variance: $s^2 = \sum(x_i - \bar{(x)})^2/(n-1)$. If mean μ is known, use n. $s^2 = \sum(x_i - \mu)^2/n$

21. Confidence intervals:

two-sided: $Prob(L < \mu < R) = 1 - \alpha \to P(\bar{X} - Z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$;

one-sided: $Prob(\mu > R) = \alpha \to P(\bar{X} - \mu/\sigma\sqrt{n} < Z_\alpha) \to \mu > \bar{X} - Z_\alpha \sigma/\sqrt{n}$

If sample is small and population variance is not known, use sample variance $s^2 = \sum(x_i - \bar{x})^2/n - 1$ and use t-distribution instead of normal distribution.

t-distribution: $f(t) = (1 + \dfrac{t^2}{n-1})^{-n/2}$, n is dof.

Then the t confidence interval on μ is $\bar{x} - t_{\alpha/2, n-1} s/\sqrt{n} < \mu < \bar{x} + t_{\alpha/2, n-1} s/\sqrt{n}$

Confidence interval on the variance: $\dfrac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} < \sigma^2 < \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$

Large sample confidence interval for a population propotion:

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

22. Hypothesis Test

	decide H_0	decide H_1
true H_0 probability	Correct action $1 - \alpha$	Type I error α
true H_1 probability	Type II error β	Correct action power $= 1 - \beta$

$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$

If H_1 is **two-sided** hypothesis, P-value is $2(1 - \Phi(|Z|))$, where $Z = (\bar(X) - \mu_0)/(s/\sqrt{n})$. If α is given, bounds are $\mu_0 \mp z_{\alpha/2} * s$ to reject null hypothesis.

For **one-sided** $\mu_1 > \mu_0$, it's $1 - \Phi(Z)$; for $\mu_1 < \mu_0$, it's $\Phi(Z)$.

If sample size n is small, use t-distribution with n-1 DOF for two-sided P-value: $2(1 - CDF_{Tdist}(|T|))$ where $T = \bar{X} - \mu_0/s\sqrt{n}$. Use $\mu_0 \mp t_{\alpha/2, n-1} T$ to **reject null hypothesis**.

Type II error and choice of **sample size**, $n = z_{\alpha/2} + z_\beta{}^2\sigma^2/\delta^2$, where $\delta = \mu - \mu_0$

m independent null hypothesis, at least one is false at significant threshold α_1: Family-Wise Error Rate=$1 - (1 - \alpha_1)^m$; to get FWER ¡ α, $\alpha_1 = \alpha/m$

Hypothesis test for a **difference** in means:

$H_0 : \mu_1 - \mu_2 = \Delta_0 = 0$, test statistic: $Z_0 = \dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

Alternative Hypotheses	*P*-Value	Rejection Criterion For for Fixed-Level Tests
H_1: $\mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
H_1: $\mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
H_1: $\mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

If $\sigma_1^2 \neq \sigma_2^2$, t-distribution with **DOF** $v = n_1 + n_2 - 2$

23. Goodness of fit test Pearson chi^2 goodness of fit test:

$\chi_0^2 = \sum_{i=1}^k (O_i - E_i)^2/E_i$, where O_i is observed number and E_i is expected number. P-value=P(H_0 is correct)= $1 - CDF_{chi-squared}(\chi_0^2, k-1)$.

How to test hypothesis if samples are drawn from same population: $P(group1; color = green) = P(group1)P(color = green)$.

$E_{green}(group1) = n_{tot}(group1/n_{tot})(green/n_{tot})$. And $\chi^2 = \sum^{n_{tot}}_{groups\&colors}(O_{color}(group) - E_{color}(group))^2/E_{color}(group)$, where DOF is (colors-1)(groups-1)

Goodness of fit with a PDF defined by m parameters

- As before: k classes (e.g. M&M colors)
- Use parameter estimators to find the best parameters for the fit
  - Method of moments
  - MLE: method of maximum likelihood
- Use chi-squared distribution with k-1-m degrees of freedom
- As before: if E_i <3, group it until E_{group} >=3 make k equal to the new number of bins

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

Confidence interval for population variance: $\chi^2_{n-1} = (n-1)S^2/\sigma^2$.

Interval form: $\dfrac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}} < \sigma^2 < \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}$

24. Regression analysis

$Y = \beta_0 + \beta_1 X + \epsilon$, $\beta_1 = Cov(X, Y)/Var(X)$; $\beta_0 = E(Y) - \beta_1 E(X)$

Use least squares to estimate: $\beta_1 = \dfrac{\sum y_i x_i - (\sum y_i)(\sum x_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$

analysis of variance $\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2 \rightarrow SS_T=SS_R+SS_E$

coefficient of determination: $R^2=SS_R/SS_T=1-SS_E/SS_T$

Estimate σ_e^2: $SS_E=\sum e_i^2=(n-2)\sigma_e^2$

Slope property: $E(\hat{\beta}_1)=\beta_1; V(\hat{\beta}_1)=\sigma^2/S_{xx}=\hat{\sigma_e^2}/n\sigma_x^2$

Intercept property: $E(\hat{\beta}_0)=\beta_0; V(\hat{\beta}_0)=\sigma^2[1/n+\bar{x}^2/S_{xx}]=\sigma_e^2[1+\mu_x^2/\sigma_x^2]/n$

Hypothesis test: H0: $\beta_1=0$; H1: $\beta_1\neq 0$

Use Z-test for large n: $Z=\hat{\beta}_1/(\hat{\sigma_e}/\sigma_x\sqrt{n})$. Reject H0 if $|Z|>Z_{\alpha/2}$

Use t-test for smaller n: $Z=\hat{\beta}_1/(\hat{\sigma_e}/\sigma_x\sqrt{n})$. Reject H0 if $|Z|>t_{\alpha/2,n-2}$

25. Multiple linear regression

$\mathbf{y}=\mathbf{X}\beta+\epsilon$, where least square is $L=\sum e_i^2=(y-X\beta)'(y-X\beta)$. $dL/d\beta=0\rightarrow\hat{\beta}=(X'X)^{-1}X'y$

Property: $E(\hat{\beta})=\beta$, Covariance Matrix: $C=(X'X)^{-1}$; $V(\hat{\beta}_j)=\sigma_e^2C_{jj}; cov(\hat{\beta}_i,\hat{\beta}_j)=\sigma_e^2C_{ij}$

Estimate σ_e^2, $\hat{\sigma_e^2}=SS_E/n-p$, where $p=k+1$

$R^2=1-SS_E/SS_T$; adjusted R-square: $R^2_{adj}=1-\dfrac{SS_E/(n-p)}{SS_T/(n-1)}$

26. Clustering algorithm

Hierarchical: agglomerative (eg, UPGMA), divisive

Non-hierarchical: PCA, K-means

- p x p symmetric matrix *R* of corr. coefficients $r_{ij}=\frac{\sigma_{ij}}{\sigma_i\sigma_j}$
- *R*=*n*$^{-1}$*Z'*Z* is a "square" of the matrix Z of standardized r.v.: $z_{\alpha k}=\frac{x_{\alpha k}-\mu_k}{\sigma_k}$ → all eigenvalues of R are non-negative
- Diagonal elements=1 → *tr(R)=p*
- Can be diagonalized: *R=V*D*V'* where *D* is the diagonal matrix
- d(1,1) –largest eig. value, d(p,p) – the smallest one
- The meaning of V(i,k) – contribution of the data type i to the k-th eigenvector
- tr(D)=p, the largest eigenvalue d(1,1) absorbs a fraction =d(1,1)/p of all correlations can be ~100%
- Scores: Y=Z*V: n x p matrix. Meaning of *Y(α,k)* – participation of the sample # α in the *k*-th eigenvector

Distances: Euclidean: $\sqrt{\sum_i(x_i-y_i)^2}$; city block (Manhattan): $\sum_i|x_i-y_i|$; Canberra: $\sum(\dfrac{x_i-y_i}{x_i+y_i})$; correlation coefficient: $1-\rho(x,y)=1-\dfrac{Cov(x,y)}{\sqrt{Var(x).Var(y)}}$