I Introduction

• Std. nor. distr. N(0,1) has cum. distr. fn.

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-y^2/2}dy$$

II Random sampling

• If we pick at random one element from the pop., then its value x is a realisation of a random var. X whose distr. is the pop. distr.

• Pop. mean $\mu=E(X)=\frac{1}{N}\sum_{i=1}^{N}x_i$ & pop. std. dev. $\sigma=\sqrt{Var(X)}$ & pop.tot $\tau=\sum_{i=1}^{N}x_i=N\mu$

• Prob. distr. curve depends on estimating μ & σ.

• S.O.R produces a simple random sample.

• S.W.R produces an I.I.D sample.

1. Point estimation

• Sampling distribution has mean $\mu_{\hat\Theta}=E(\hat\Theta)$ and var. $\sigma^2_{\hat\Theta}=E(\hat\Theta-\mu_{\hat\Theta})^2$.

2. Sample mean and sample variance

• Sample mean $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ & sample std. dev. $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{n-1}(\overline{x^2}-\bar{x}^2)$

• Independent and identically distributed(I.I.D) sample: sample mean $\bar{x}$ and sample var s^2 are unbiased estimators for the pop. mean μ and var σ^2 $E(\bar{X})=\mu, Var(\bar{X})=\frac{\sigma^2}{n}, E(S^2)=\sigma^2, Var(S^2)=\frac{\sigma^4}{n}(E(\frac{X-\mu}{\sigma})^4-\frac{n-3}{n-1})$

• Dichotomous case: sample mean turns into a sample proportion $\hat{p}=\bar{x}$ giving an unbiased estimate of p $\mu=p, \sigma^2=p(1-p)$

• Estimated standard errors for the sample mean and proportion $s_{\bar{x}}=\frac{s}{\sqrt{n}}$ & $s_{\hat{p}}=\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$

• Simple random sampling(S.R.S): sample mean $\bar{x}$ is an unbiased estimate for pop. mean μ & sample var. s^2 is a biased estimate of σ^2 $E(S^2)=\sigma^2\frac{N}{N-1}$ & $Var(\bar{X})=\frac{\sigma^2}{n}(1-\frac{n-1}{N-1})$. Unbiased estimate of $Var(\bar{X})$ is $s^2_{\bar{x}}=\frac{s^2}{n}(1-\frac{n}{N})$

• Sampling without replacement(S.O.R): estimated standard errors $s_{\bar{x}}=\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$ & $s_{\hat{p}}=\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{n}{N}}$

3. Approximate confidence intervals

• By the Central Limit Theorem, the sample mean distr. is approx. normal $\bar{X}\approx N(\mu,\frac{\sigma^2}{n})$ in that for large sample sizes n, we have $P(|\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}|>z)\approx 2(1-\phi(z))$. Since $S_{\bar{X}}\approx\sigma_{\bar{X}}$, we have $P(\bar{X}-zS_{\bar{X}}<\mu<\bar{X}+zS_{\bar{X}})=P(|\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}|>z)\approx 2(1-\phi(z))$. This yields the formula of 100(1–α)% 2-sided c.i for μ & p: $I_\mu=\bar{x}\pm z_{\alpha/2}s_{\bar{x}}, I_p=\hat{p}\pm z_{\alpha/2}s_{\hat{p}}$.

100(1–α)	68	80	90	95	99	99.7
$z_{\alpha/2}$	1.00	1.28	1.64	1.96	2.58	3.00

• 100(1–α)% c.i for μ means 100(1–α)% are expected to cover the true value of μ.

• Higher the confidence level, the wider is confidence interval. Larger the sample, the narrower is confidence interval.

4. Stratified random sampling

• Strat. prop. for each strat. j, j=1,$\cdots$,k is $\omega_j=\frac{N_j}{N}$ s.t. $\omega_1+\cdots+\omega_k=1$

• Pop. mean is $\mu=\omega_1\mu_1+\cdots+\omega_k\mu_k$ & pop. var. is $\sigma^2=\overline{\sigma^2}+\sum_{j=1}^{k}\omega_j(\mu_j-\mu)^2$

• Stratified sample mean is $\bar{x}_s=\omega_1\bar{x}_1+\cdots+\omega_k\bar{x}_k$ where $\bar{x}_j$ = est. of sample mean of sample size n_j for each strat. j

• Var. of $\bar{X}_s$ is $Var(\bar{X}_s)=\sigma^2_{\bar{X}_s}=\frac{\omega_1^2\sigma_1^2}{n_1}+\cdots+\frac{\omega_k^2\sigma_k^2}{n_k}$

• Opt. alloc. is $n_j=\frac{n\omega_j\sigma_j}{\bar\sigma}$ where σ_j = sample std. dev. of sample size n_j for each strat. j & $\bar\sigma=\omega_1\sigma_1+\cdots+\omega_k\sigma_k$ & n=tot. sample size

• Prop. alloc. is $n_j=n\omega_j$

• Var. of est. of pop. mean obtd using s.r.s is $Var(\bar{X})=\frac{\sigma^2}{n}$

• Var. of est. of pop. mean obtd using prop. alloc is $Var(\bar{X}_{sp})=\frac{\overline{\sigma^2}}{n}$ where $\overline{\sigma^2}=\omega_1\sigma_1^2+\cdots+\omega_k\sigma_k^2$

• Var. of est. of pop. mean obtd using opt. alloc. is $Var(\bar{X}_{so})=\frac{\bar{\sigma}^2}{n}$

• $Var(\bar{X})\geq Var(\bar{X}_{sp})\geq Var(\bar{X}_{so})$

• Approx. c.i. $I_\mu=\bar{x}_s\pm z_{\alpha/2}s_{\bar{x}_s}$

III Parameter estimation

1. Method of moments

Suppose an iid-sample from a pop. distr. is characterised by a pair of parameters (θ_1,θ_2). Suppose we have the foll. formulas for the 1st and 2nd pop. moments: $E(X)=f(\theta_1,\theta_2)$, $E(X^2)=g(\theta_1,\theta_2)$. M.O.M est. $(\tilde\theta_1,\tilde\theta_2)$ are found after replacing the pop. moments with the corr. sample moments, and solving the obtained eqns $\bar{x}=f(\tilde\theta_1,\tilde\theta_2)$, $\bar{x^2}=g(\tilde\theta_1,\tilde\theta_2)$. This approach is justified by the Law of Large Numbers $\frac{X_1+\cdots+X_n}{n}\to\mu$, $\frac{X_1^2+\cdots+X_n^2}{n}\to E(X^2)$, $n\to\infty$.

• Geometric model- m.o.m est. for para. $\theta=p$ is $\bar{x}=\frac{1}{\tilde{p}}$, approx. c.i. for p is $I_p=\frac{1}{\bar{x}\pm z_{\alpha/2}s_{\bar{x}}}$, obs. freq. given, exptd. freq.= $n(1-\tilde{p})^{j-1}\tilde{p}$

2. Maximum likelihood estimation

• In a parametric setting, given a parameter value θ, the observed sample $(x_1,\cdots,x_n)$ is a realisation of the random vector $(X_1,\cdots,X_n)$ which has a certain joint distr. $f(y_1,\cdots,y_n|\theta)$ as a fn. of possible values $(y_1,\cdots,y_n)$. Fixing the variables $(y_1,\cdots,y_n)=(x_1,\cdots,x_n)$ and allowing the parameter value θ to vary, we obtain the likelihood fn. $L(\theta)=f(x_1,...x_n|\theta)$. The maximum likelihood estimate $\hat\theta$ of θ is the value of θ that maximises $L(\theta)$.

• Example: binomial model- n obs. $X\sim Bin(n,p)$. From $\mu=np$, m.o.m est. $\tilde{p}=\frac{x}{n}$ is the sample proportion. Likelihood fn.is $L(p)=\binom{n}{x}p^x(1-p)^{n-x}$. Log-likelihood function is $l(p)=\ln L(p)$. Then $l\prime(p)=0$ gives MLE of pop. prop. is the sample prop. $\hat{p}=\frac{x}{n}$

3. Sufficiency

• Example: Bernoulli distr.-For 1 Ber. trial, $f(x)=P(X=x)=p^x(1-p)^{1-x}$, $x\in\{0,1\}$ and for n Ber. trials, $f(x_1,\cdots,x_n|p)=\Pi_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{n\bar{x}}(1-p)^{n-n\bar{x}}$. Thus for Ber. model, no. of successes $t=x_1+\cdots+x_n=n\bar{x}$ is a suff. statistic whose distr. is $T\sim Bin(n,p)$.

• Example: Normal distr.- Nor. distr. model $N(\mu,\sigma^2)$ has 2-D sufficient statistic (t_1,t_2), $t_1=\sum_{i=1}^{n}x_i$, $t_2=\sum_{i=1}^{n}x_i^2$ which follows from $L(\mu,\sigma)=\Pi_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(1/\sigma^n(2\pi)^{n/2})e^{-\frac{t_2-2\mu t_1+n\mu^2}{2\sigma^2}}$

4. Large sample prop. of the m.l.e

• Normal approx. $\hat\Theta\approx N(\theta,\frac{1}{nI(\theta)})$. Let $g(x,\theta)=\frac{\partial^2\ln f(x|\theta)}{\partial\theta^2}$ & Fisher information $I(\theta)=-E[g(X,\theta)]=-\int g(x,\theta)f(x|\theta)dx$. Approx. 100(1-$\alpha$)% c.i. $I_\theta=\hat\theta\pm z_{\frac{\alpha}{2}}\frac{1}{\sqrt{nI(\theta)}}$

• Example: exponential model- From $\mu=\frac{1}{\theta}$, we find $\tilde\theta=\frac{1}{\bar{x}}$. Likelihood fn. $L(\theta)=\theta^n e^{-\theta(x_1+\cdots+x_n)}$. For exp. model, $t=x_1+\cdots+x_n$ is a suff. statistic & M.L.E is $\hat\theta=\frac{1}{\bar{x}}$

5. Gamma distribution

• Den. fn of Gamma distr. is $f(x|\alpha,\lambda)=\frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}, 0\leq x<\infty$ where shape parameter $\alpha>0$ and scale parameter $\lambda>0$. $(t_1,t_2)=(x_1+\cdots+x_n, x_1\cdots x_n)$ is a pair of sufficient statistics. Likelihood fn. is $L(\alpha,\lambda)=\prod_{i=1}^{n}\frac{1}{\Gamma(\alpha)}\lambda^\alpha x_i^{\alpha-1}e^{-\lambda x_i}$. Log likelihood of i.i.d sample is $l(\alpha,\lambda)=\sum_{i=1}^{n}[\alpha\ln\lambda+(\alpha-1)\ln x_i-\lambda x_i-\ln\Gamma(\alpha)]=n\alpha\ln\lambda+(\alpha-1)\sum_{i=1}^{n}\ln x_i-\lambda\sum_{i=1}^{n}x_i-n\ln\Gamma(\alpha)$.

Partial der. are $\frac{\partial l}{\partial\alpha}=n\ln\lambda+\sum_{i=1}^{n}\ln x_i-n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$, $\frac{\partial l}{\partial\lambda}=\frac{n\alpha}{\lambda}-\sum_{i=1}^{n}x_i$. Setting these to 0, $\hat\lambda=\frac{\hat\alpha}{\bar{x}}$, $n\ln\hat\alpha-n\ln\bar{x}+\sum_{i=1}^{n}\ln x_i-n\frac{\Gamma'(\hat\alpha)}{\Gamma(\hat\alpha)}=0$.Eqns. cannot be solved in closed form and $\therefore$ we use iter. meth. for finding the roots where the intial val. obtd. by m.o.m est.

• Parametric bootstrap- For initial values, we apply m.o.m est. formula are $\hat\alpha=\frac{\hat\mu^2}{\hat\sigma^2}$ & $\hat\lambda=\frac{\hat\mu}{\hat\sigma^2}$, $\hat\mu=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ & $\hat\sigma^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2$. If we could simulate from the true population distribution $Gam(\alpha,\lambda)$, then B samples of size n would generate B independent estimates $\hat\alpha_j$. Then the standard deviation of the sampling distribution would give us the desired std. err. $s_{\hat\alpha}=\sqrt{\frac{1}{B}\sum_{j=1}^{B}(\hat\alpha_j-\bar\alpha)^2}, \bar\alpha=\frac{1}{B}\sum_{j=1}^{B}\hat\alpha_j$

• Diff. betwn. parametric & nonparametric bootstrap- In parametric bootstrap, we have a known (assumed) distr. $l(\theta)$ with unknown parameter θ. We estimate θ by $\hat\theta$ and draw samples from the distr. $l(\hat\theta)$. In non-parametric bootstrap, we do not assume an underlying distr and instead resample from the set of original samples $x_1,\cdots,x_n$.

• Bootstrap is a resampling technique used to study the sampling distribution of a parameter estimator. In the parametric bootstrap resampling is done from the given parametric distribution with the unknown parameters replaced by their estimates obtained from the underlying sample. In the non-parametric bootstrap resampling is performed with replacement directly from the the underlying sample.

6. Exact confidence intervals

• t-distribution: $\frac{\bar{X}-\mu}{S_{\bar{X}}}\sim t_{n-1}$

• Exact 100 (1-α)% c.i. for mean: $I_\mu=\bar{x}\pm t_{n-1}(\alpha/2)s_{\bar{x}}$

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

• Exact 100 (1-α)% c.i. for var: $I_{\sigma^2}=(\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)},\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)})$

IV Hypothesis testing

1. Statistical significance

• A rule based on data for choosing between 2 mutually exclusive hypotheses
null hypothesis H_0: effect of interest is zero,
alternative H_1: effect of interest is not zero.
A decision rule for hypotheses testing is based on a test statistic $t=t(x_1,\cdots,x_n)$, a fn. of the data with distinct typical values under H_0 & H_1. The task is to find an appropriately chosen rejection region R and reject H_0 in favor of H_1 if and only if $t\in$ R.

• Four imp condn. prob.-
significance level/type I error-$\alpha=P(T\in R|H_0)$,
specificity of the test-$1-\alpha=P(T\notin R|H_0)$,
type II error-$\beta=P(T\notin R|H_1)$,
sensitivity or power-$1-\beta=P(T\in R|H_1)$

• In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a false positive outcome), while a type II error is the failure to reject a false null hypothesis (also known as a false negative outcome).

• A significance level is a pre-decided limit for when we reject the null hypothesis.

• A p-value is the probability of obtaining a test statistic value as extreme or more extreme than the observed one, given that H_0 is true. For given α, reject H_0, if p-value $\leq\alpha$, and do not reject H_0, if p-value $>\alpha$. Observe that the p-value depends on the data and therefore, is a realisation of a random variable P. The source of randomness is in the sampling procedure: if you take another sample, you obtain a different p-value.To illustrate, suppose we are testing $H_0:\theta=\theta_0$ vs. $H_1:\theta>\theta_0$. Suppose the null hypothesis is true. Given $z_{obs}=z$, p-value is $p=P(Z>z)=1-\phi(z)$ and in terms of the random variables $P=P(Z>Z_{obs})=1-\phi(Z_{obs})$.

• Under H_0, $P(P>p)=P(1-\phi(Z_{obs})>1-\phi(z))=P(\phi(Z_{obs})<\phi(z))=P(Z_{obs}<z)=\phi(z)=1-p$. $\therefore$ P-value has uniform null distr.

2. Large-sample test for the proportion

Bin model X ∼ Bin(n,p). sample prop. $\hat{p}=\frac{x}{n}$ test statistic $z=\frac{x-np_0}{\sqrt{np_0(1-p_0)}}=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. 3 diff. alternative hypotheses: 1-sided H_1: $p>p_0$, 1-sided H_1: $p<p_0$, 2-sided H_1: $p\neq p_0$.

Alter. H_1	Rej. rule	P-value
$p>p_0$	$z\geq z_\alpha$	$P(Z\geq z_{obs})$
$p<p_0$	$z\leq -z_\alpha$	$P(Z\leq z_{obs})$
$p\neq p_0$	$z\leq -z_{\alpha/2}$ or $z\geq z_{\alpha/2}$	$2P(Z\geq\vert z_{obs}\vert)$

• Power function- Consider two simple hypotheses H_0: $p=p_0$ & H_1: $p=p_1$ assuming $p_1>p_0$. The null distribution of Y is approximately normally distributed with parameters $(np_0,np_0(1-p_0))$. At the significance level α, the rejection region for the one-sided alternative is $\frac{X-np_0}{\sqrt{np_0(1-p_0)}}\geq z_\alpha$. The power function of the one-sided test can be computed using the normal approximation for $\frac{X-np_1}{\sqrt{np_1(1-p_1)}}$ under H_1: $Pw(p_1)=P(\frac{X-np_0}{\sqrt{np_0(1-p_0)}}\geq z_\alpha|H_1)=P(\frac{X-np_1}{\sqrt{np_1(1-p_1)}}\geq\frac{z_\alpha\sqrt{p_0(1-p_0)}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1(1-p_1)}}|H_1)\approx 1-\phi(\frac{z_\alpha\sqrt{p_0(1-p_0)}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1(1-p_1)}})$, Now, since under the alternative hypothesis X is approximately normally distributed with parameters (np_1,np_1q_1), we get $\beta\approx\phi(\frac{z_\alpha\sqrt{p_0(1-p_0)}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1(1-p_1)}})$. This leads to the equation $\frac{z_\alpha\sqrt{p_0(1-p_0)}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1(1-p_1)}}=-z_\beta$ which gives the formula for sample size - $\sqrt{n}=\frac{z_\alpha\sqrt{p_0(1-p_0)}+z_\beta\sqrt{p_1(1-p_1)}}{|p_1-p_0|}$

• If the alternatives are very close to each other, the denominator tends to zero and hence the sample size becomes very large.

• If we decrease the levels α and β, the values z_α and z_β from the normal distribution table become larger and the sample size will be larger as well. If we want to have more control over both types of errors, we have to collect more data.

3. Small-sample test for the proportion

• For small n, we use exact null distribution $X\sim Bin(n,p_0)$. $P(X\geq x)=\sum_{j=x}^{n}\binom{n}{j}p_0^j(1-p_0)^{n-j}$

4. Two tests for the mean

• Large-sample test for mean- pop. distr. is not necessarily normal- sample size n is sufficiently large-compute the rejection region using an approximate null distr. $T\approx^{H_0}N(0,1)$

• One-sample t-test-pop. distr. is normal-small n-compute the rejection region using an exact null distr. $T\sim^{H_0}t_{n-1}$

• C.I. method of hypotheses testing- at sig. level α, rejection rule is $R=\{\mu_0\notin I_\mu\}$. Reject H_0: $\mu=\mu_0$ if the interval does not cover val. of μ_0.

5. Likelihood ratio test

For testing $H_0:\theta=\theta_0$ against $H_1:\theta=\theta_1$, we use likelihood ratio as test statistic-$\Lambda=\frac{L(\theta_0)}{L(\theta_1)}$. Large values of Λ suggest that H_0 explains the data set better than H_1, while small Λ indicates that H_1 explains the data set better. Likelihood ratio test rejects H_0 for small values of Λ.

• Neyman-Pearson lemma: the likelihood ratio test is optimal in the case of two simple hypothesis.

6. Pearson's chi-squared test/Goodness of fit chi-square test

• Suppose that each of n indep. obs. belongs to one of J classes with prob. $(p_1,\cdots,p_J)$. Such data are summarised as the vector of observed counts whose joint distribution is multinomial $(O_1,\cdots,O_J)\sim Mn(n;p_1,\cdots,p_J)$, $P(O_1=k_1,\cdots,O_J=k_J)=\frac{n!}{k_1!\cdots k_J!}p_1^{k_1}\cdots p_J^{k_J}$. Consider a parametric model for the data $H_0:(p_1,\cdots,p_J)=(v_1(\lambda),\cdots,v_J(\lambda))$ with unknown parameters $\lambda=(\lambda_1,\cdots,\lambda_r)$. To see if the proposed model fits the data, compute $\hat\lambda$, the m.l.e of λ, and then expected cell counts $E_j=nv_j(\hat\lambda)$. Chi-squared test statistic $\chi^2=\sum_{j=1}^{J}\frac{(O_j-E_j)^2}{E_j}$ whose approx. null distr. is χ^2_{J-1-r} where J = no. of cells & r = no. of indep. para. estimtd. from data.

• Example: geometric model

7. Example: sex ratio

• Simple hypothesis $H_0:p_j=\binom{n}{j}2^n$. $E_j=Np_j$

• Comp. hyp. $H_0:p_j=\binom{n}{j}\hat{p}^j(1-\hat{p})^{n-j}$. $E_j=Np_j$

V Bayesian inference

• Posterior distr. $h(\theta|x)$ using the Bayes formula (Bayes Probability Law) $h(\theta|x)=\frac{f(x|\theta)g(\theta)}{\phi(x)}$ where $\phi(x)=\int f(x|\theta)g(\theta)d\theta$ or $\sum_\theta f(x|\theta)g(\theta)$.

Treating ϕ as a constant and the Bayes formula can be summarised as $posterior \propto likelihood \times prior$.

1. Conjugate priors

• Beta distribution- $f(p)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}p^{b-1}, 0<p<1$ with mean and variance $\mu=\frac{a}{a+b}, \sigma^2=\frac{\mu(1-\mu)}{a+b+1}$

• Dirichlet distribution- Density fn. $f(p_1,\cdots,p_r)=\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_r)}p_1^{\alpha_1-1}\cdots p_r^{\alpha_r-1}$

• List of Conjugate priors

Data distr.	Prior	Posterior distr.
$X_1,\cdots,X_n \sim N(\mu,\sigma^2)$	$\mu\sim N(\mu_0,\sigma_0^2)$	$N(\gamma_n\mu_0+(1-\gamma_n)\bar{x};\gamma_n\sigma_0^2)$
$X\sim Bin(n,p)$	$p \sim Beta(a,b)$	$Beta(a+x,b+n-x)$
$(X_1,\cdots,X_r)\sim Mn(n;p_1,\cdots,p_r)$	$(p_1,\cdots,p_r)\sim Dir(\alpha_1,\cdots,\alpha_r)$	$Dir(\alpha_1+x_1,\cdots,\alpha_r+x_r)$
$X_1,\cdots,X_n\sim Geom(p)$	$p\sim Beta(a,b)$	$Beta(a+n,b+n\bar{x}-n)$
$X_1,\cdots,X_n \sim Pois(\mu)$	$\mu\sim Gam(\alpha_0,\lambda_0)$	$Gam(\alpha_0+n\bar{x},\lambda_0+n)$
$X_1,\cdots,X_n\sim Gam(\alpha,\lambda)$	$\lambda\sim Gam(\alpha_0,\lambda_0)$	$Gam(\alpha_0+\alpha n,\lambda_0+n\bar{x})$

• posterior pseudo-counts = prior pseudo-counts plus sample counts

• Normal-Normal model Shrinkage factor- $\gamma_n=\frac{\sigma^2}{\sigma^2+n\sigma_0^2}$

• Binomial-Beta model-
Simple demonstration that beta distribution gives a conjugate prior to the binomial likelihood.
$prior\propto p^{a-1}(1-p)^{b-1}$
$likelihood\propto p^x(1-p)^{n-x}$
$posterior \propto prior\times likelihood \propto p^{a+x-1}(1-p)^{b+n-x-1}$.

2. Bayesian estimation

• In terms of decision theory, we are looking for an optimal action $a=\{assign value to unknown parameter \theta\}$. The optimal a depends on the choice of the loss function $l(\theta,a)$. Bayes action minimises posterior risk $R(a|x)=E(l(\Theta,a)|x)$ so that $R(a|x)=\int l(\theta,a)h(\theta|x)d\theta$ or $R(a|x)=\sum_\theta l(\theta,a)h(\theta|x)$. There are 2 loss fns leading to two Bayesian estimators.

1. Zero-one loss fn and max a posteriori probability
Zero-one loss fn: $l(\theta,a)=1_{\theta\neq a}$
Using zero-one loss fn., the posterior risk is $R(a|x)=\sum_{\theta\neq a}h(\theta|x)=1-h(a|x)$. It follows that to minimise the risk we have to maximise the posterior probability. We define $\hat\theta_{map}$ as the value of θ that maximises $h(\theta|x)$. Observe that with the uninformative prior, $\hat\theta_{map}=\hat\theta_{mle}$.

2. Squared error loss function and posterior mean estimate
Squared error loss: $l(\theta,a)=(\theta-a)^2$. Using squared error loss function, the posterior risk is $R(a|x)=E((\Theta-a)^2|x)=Var(\Theta|x)+[E(\Theta|x)-a]^2$. Since the first component is independent of a, we minimise the posterior risk by putting $\hat\theta_{pme}=E(\Theta|x)$.

• Multinomial Dirichlet- 2 Bayesian estimates-
1. Prior- $\hat\theta_{map}=\hat\theta_{mle}=(\frac{\alpha_1}{\alpha_1+\cdots+\alpha_r},\cdots,\frac{\alpha_r}{\alpha_1+\cdots+\alpha_r})$
2. Posterior mean estimate-$\hat\theta_{pme}=(\frac{\alpha_1+x_1}{\alpha_1+x_1+\cdots+\alpha_r+x_r},\cdots,\frac{\alpha_r+x_r}{\alpha_1+x_1+\cdots+\alpha_r+x_r})$

3. Credibility interval

• Let x be the data. For a confidence interval formula $I_\theta=(a_1(x),a_2(x))$, the parameter θ is an unknown constant and a confidence interval is random $P(a_1(X)<\Theta<a_2(X))=1-\alpha$. A credibility interval $J_\theta=(b_1(x),b_2(x))$ is treated as a nonrandom interval, while θ is generated by the posterior distribution of a random variable Θ. A credibility interval is computed from the posterior distribution $P(b_1(x)<\Theta<b_2(x)|x)=1-\alpha$.

4. Bayesian hypotheses testing

We consider the case of two simple hypotheses. Choose between $H_0:\theta=\theta_0$ & $H_1:\theta=\theta_1$ using not only the likelihoods of the data $f(x|\theta_0), f(x|\theta_1)$ but also prior probabilities $P(H_0)=\pi_0$, $P(H_1)=\pi_1$. In terms of the rejection region R the decision should be taken depending of a cost function. c_0 is the error type I cost and c_1 is the error type II cost. For a given set R, the average cost is the weighted mean of two values c_0 and c_1 is $c_0\pi_0P(X\in R|H_0)+c_1\pi_1P(X\notin R|H_1)=c_1\pi_1+\int_R(c_0\pi_0f(x|\theta_0)-c_1\pi_1f(x|\theta_1))dx$. It follows that the rejection region minimising the average cost is $R=\{x: c_0\pi_0f(x|\theta_0)<c_1\pi_1f(x|\theta_1)\}$. Thus the optimal decision rule becomes to reject H_0 for small values of the likelihood ratio when $\frac{f(x|\theta_0)}{f(x|\theta_1)}<\frac{c_1\pi_1}{c_0\pi_0}$ or for small posterior odds, $\frac{h(\theta_0|x)}{h(\theta_1|x)}<\frac{c_1}{c_0}$.

VI Summarising data

1. Empirical probability distribution

Empirical distr. fn. $\hat F(x)=\frac1n\sum_{i=1}^n 1_{\{x_i\le x\}}$.
If the data describes life lengths, then it is more convenient to use the empirical survival fn. $\hat S(x)=1-\hat F(x)$, the proportion of the data greater than x. If the life length T has distr. fn. $F(t)=P(T\le t)$, then its survival function is $S(t)=P(T>t)=1-F(t)$. Hazard function $h(t)=\frac{f(t)}{S(t)}$ where $f(t)=F\prime(t)$ is the probability density fn. The hazard fn. (also known as the failure rate, hazard rate, or force of mortality) is the ratio of the probability density function to the survival fn. The hazard function is the mortality rate at age t: $P(t<T\le t+\delta|T\ge t)=\frac{P(t<T\le t+\delta)}{P(T\ge t)}=\frac{F(t+\delta)-F(t)}{S(t)}\sim\delta h(t),\delta\to0$. The hazard function can be viewed as the negative of the slope of the log survival fn: $h(t)=-\frac{d}{dt}\ln S(t)=-\frac{d}{dt}\ln(1-F(t))$. A constant hazard rate $h(t)=\lambda$ corresponds to the exponential distribution $Exp(\lambda)$.

2. Density estimation

3. Quantiles and QQ-plots

For a given distr. F and $0\le p\le1$, the p-quantile is $x_p=Q(p)$. $x_{(k)}$ is called the empirical $(\frac{k-0.5}{n})$quantile.. QQ-plot is a scatter plot of n dots with coordinates $(x_{(k)},y_{(k)})$.

4. Testing normality

• Coefficient of skewness: $\beta_1=\frac{E(X-\mu)^3}{\sigma^3}$, sample skewness: $b_1=\frac{1}{s^3n}\sum_{i=1}^n(x_i-\bar x)^3$. Depending on the sign of the coefficient of skewness with distinguish between symmetric $\beta_1=0$, skewed to the right $\beta_1>0$, and skewed to the left $\beta_1<0$ distr.

• Kurtosis $\beta_2=\frac{E(X-\mu)^4}{\sigma^4}$, sample kurtosis: $b_2=\frac{1}{s^4n}\sum_{i=1}^n(x_i-\bar x)^4$. Kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable, although some sources are insistent that heavy tails, and not peakedness, is what is really being measured by Kurtosis. For the normal distribution, kurtosis coefficient takes value $\beta_2=3$. Leptokurtic distribution: $\beta_2>3$ (heavy tails). Platykurtic distr: $\beta_2<3$ (light tails).

5. Measures of location

• Sample median $\hat m=x_{(k)}$, if $n=2k-1$, and $\hat m=\frac{x_{(k)}+x_{(k+1)}}{2}$, if $n=2k$.
$I_m=(x_{(k)},x_{(n-k+1)})$ is a $100p_k$% c.i. for the pop. median m.

• sign test The sign test is a non-parametric test of $H_0: m=m_0$ against the two-sided alternative $H_0: m\neq m_0$. The sign test statistic $y_0=\sum_{i=1}^n 1_{\{x_i\le m_0\}}$ counts the number of observations below the null hypothesis value. It has a simple null distribution $Y_0\overset{H_0}{\sim}Bin(n,0.5)$. Connection to the above c.i formula: reject H_0 if m_0 falls outside the corresponding c.i. $I_m=(x_{(k)},x_{(n-k+1)})$.

• trimmed means- α-trimmed mean $\bar x_\alpha$ = sample mean without $\frac{n\alpha}{2}$ smallest and $\frac{n\alpha}{2}$ largest observations.

• Nonparametric bootstrap- Substitute the population distribution by the empirical distribution. Then a bootstrap sample is obtained by resampling with replacement from the original sample $(x_1,\cdots,x_n)$. Generate many bootstrap samples of size n to approximate the sampling distribution for an estimator like trimmed mean, sample median, or s.
The difference between non parametric bootstrap and parametric bootstrap is that parametric is with the normality assumption, and non parametric is without the normality assumption.

6. Measures of dispersion

VII Comparing two samples

• We wish to compare 2 pop. distr. with means and std. dev. (μ_1,σ_1), (μ_2,σ_2) based on 2 iid-samples $(x_1,\cdots,x_n)$ and $(y_1,\cdots,y_m)$ from these 2 pop. Two sample means $\bar x$, $\bar y$ and their std errors $s_{\bar x}=\frac{s_1}{\sqrt n}, s_1^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)^2$ & $s_{\bar y}=\frac{s_2}{\sqrt m}, s_2^2=\frac{1}{m-1}\sum_{i=1}^m(y_i-\bar y)^2$. Diff. $\bar x-\bar y$ is an unbiased est. of $\mu_1-\mu_2$. We are interested in finding the std. error of $\bar x-\bar y$ & an interval est. for $\mu_1-\mu_2$ & testing the null hypothesis of equality $H_0:\mu_1=\mu_2$.

1. Two indep. samples: comparing pop. means

If $(X_1,\cdots,X_n)$ is indep. from $(Y_1,\cdots,Y_m)$, then $Var(\bar X-\bar Y)=\sigma_{\bar X}^2+\sigma_{\bar Y}^2=\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}$ & $s_{\bar x-\bar y}^2=s_{\bar x}^2+s_{\bar y}^2=\frac{s_1^2}{n}+\frac{s_2^2}{m}$ gives an unbiased estimate of $Var(\bar X-\bar Y)$.

• Large sample test for the difference between two means- If n and m are large, we can use a normal approximation $\frac{\bar X-\bar Y-(\mu_1-\mu_2)}{\sqrt{S_{\bar X}^2+S_{\bar Y}^2}}\approx N(0,1)$.
The hypothesis $H_0:\mu_1=\mu_2$ is tested using the test statistic $z=\frac{\bar x-\bar y}{\sqrt{s_{\bar x}^2+s_{\bar y}^2}}$ whose null distribution is approximated by the standard normal N(0,1).
Approximate confidence interval formula- $I_{\mu_1-\mu_2}=\bar x-\bar y\pm z_{\alpha/2}\sqrt{s_{\bar x}^2+s_{\bar y}^2}$

• Two-sample t-test The key assumption of the two-sample t-test: two normal pop. distr. $X\sim N(\mu_1,\sigma^2)$, $Y\sim N(\mu_2,\sigma^2)$ have equal variances. Given $\sigma_1^2=\sigma_2^2=\sigma^2$, the pooled sample variance $s_p^2=\frac{\sum_{i=1}^n(x_i-\bar x)^2+\sum_{i=1}^m(y_i-\bar y)^2}{n+m-2}=\frac{n-1}{n+m-2}s_1^2+\frac{m-1}{n+m-2}s_2^2$ is an unbiased estimate of the variance with $E(S_p^2)=\frac{n-1}{n+m-2}E(S_1^2)+\frac{m-1}{n+m-2}E(S_2^2)=\sigma^2$. In the equal variance two sample setting, the variance $Var(\bar X-\bar Y)=\sigma^2\frac{n+m}{nm}$ has the following unbiased estimate $s_{\bar x-\bar y}^2=s_p^2\frac{n+m}{nm}$.
Exact distribution $\frac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{S_p}\sqrt{\frac{nm}{n+m}}\sim t_{n+m-2}$. Exact confidence interval formula $I_{\mu_1-\mu_2}=\bar x-\bar y\pm t_{n+m-2}(\alpha/2)s_p\sqrt{\frac{n+m}{nm}}$. Two sample t-test uses the test statistic $t=\frac{\bar x-\bar y}{s_p}\sqrt{\frac{nm}{n+m}}$ for testing H_0:$\mu_1=\mu_2$. The null distribution of the test statistic is $T\sim t_{n+m-2}$.

• Rank sum test-It is a nonparametric test for two indep. samples, which does not assume normality of pop. distr. Assume continuous population distributions F_1 and F_2, and consider H_0: $F_1=F_2$ against H_1: $F_1\neq F_2$. The rank sum test procedure: pool the samples and replace the data values by their ranks $1,2,\cdots,n+m$ starting from the smallest sample value to the largest, and then compute two test statistics r_1 = sum of x-ranks, and r_2 = sum of y-ranks. Clearly $r_1+r_2=1+2+\cdots+(n+m)=\frac{(n+m)(n+m+1)}{2}$. The null distr. for R_1 and R_2 depend only on the sample sizes n and m. For $n\ge10$, $m\ge10$, apply the normal approximation for the null distr. of R_1 and R_2 with $E(R_1)=\frac{n(n+m+1)}{2}$, $E(R_2)=\frac{m(n+m+1)}{2}$, $Var(R_1)=Var(R_2)=\frac{mn(n+m+1)}{12}$.

2. Two indep. samples: comparing population proportion

For $X\sim Bin(n,p_1)$, $Y\sim Bin(m,p_2)$, unbiased est. of p_1 & p_2 are $\hat p_1=\frac xn$ & $\hat p_2=\frac ym$ which have standard errors $s_{\hat p_1}=\sqrt{\frac{\hat p_1(1-\hat p_1)}{n-1}}$ & $s_{\hat p_2}=\sqrt{\frac{\hat p_2(1-\hat p_2)}{m-1}}$
Large sample test for two proportions- If the samples sizes m and n are large, then an approx. c.i for p_1-p_2 is $I_{p_1-p_2}=\hat p_1-\hat p_2\pm1.96\sqrt{\frac{\hat p_1(1-\hat p_1)}{n-1}+\frac{\hat p_2(1-\hat p_2)}{m-1}}$.
We can test the null hypothesis of equality $H_0: p_1=p_2$

3. Paired samples

• Fisher's exact test-Fisher's exact test deals with the null hypothesis $H_0: p_1=p_2$ when the sample sizes m and n are not sufficiently large for applying normal approximations for the binomial distr.

	Sample 1	Sample 2	Total
Number of successes	x	y	Np=x+y
Number of failures	n-x	m-y	Nq=n+m-x-y
Sample sizes	n	m	N=n+m

Fisher's idea for this case, was to use X as a test statistic conditionally on the total number of successes x + y. Under the null hypothesis, the conditional distr. of X is hypergeometric $X\sim Hg(N,n,p)$ with parameters (N,n,p) defined by $N=n+m$, $p=\frac{x+y}{N}$. This is a discrete distr. with prob. mass fn. $P(X=x)=\frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$, $max(0,n-Nq)\le x\le min(n,Np)$. This null distr. should be used for determining the rejection rule of the Fisher test.

• Signed rank test
The sign test disregards a lot of information in the data taking into account only the sign of the differences. The signed rank test pays attention to sizes of positive and negative differences. This is a non-parametric test for the null hypothesis of no diff. H_0 : distr. of D is symmetric about its median $m=0$. The null hypothesis consists of two parts: symmetry of the distr. and $m=0$. Test statistics: either $w_+=\sum_{i=1}^n rank(|d_i|).1_{d_i>0}$ or $w_-=\sum_{i=1}^n rank(|d_i|).1_{d_i<0}$. Assuming no ties, that is $d_i\neq0$, we get $w_++w_-=\frac{n(n+1)}{2}$. The null distributions of W_+ & W_- are the same and tabulated for smaller values of n. For $n\ge20$, one can use the normal approximation of the null distr. with mean and var. $\mu_W=\frac{n(n+1)}{4}$ & $\sigma_W^2=\frac{n(n+1)(2n+1)}{24}$. The signed rank test uses more data information than the sign test but requires symmetric distribution of differences.

4. External and confounding factors

placebo effect

VIII Analysis of variance

	1-way ANOVA	2-way ANOVA
Defn.	A test that allows one to make comparisons between the means of 3 or more groups of data.	A test that allows one to make comparisons between the means of 3 or more groups of data, where two independent variables are considered.
No. of indep. var.	One.	Two.
What is being Compared?	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two indep. variables on a dependent variable and on each other.
No. of groups of samples	Three or more.	Each variable should have multiple samples.

1. One-way layout

$Y=\mu(X)+\epsilon, E(\epsilon)=0$
Normal theory model
$Y_{ij}=\mu+\alpha_i+\epsilon_{ij},\sum_i\alpha_i=0,\epsilon_{ij}\sim N(0,\sigma^2)$. Using the maximum likelihood approach, the point estimates are $\hat\mu=\bar y_{..},\hat\mu_i=\bar y_{i.},\hat\alpha_i=\bar y_{i.}-\bar y_{..}$ where $\bar y_{i.}=\frac1J\sum_j y_{ij}$ & $\bar y_{..}=\frac1I\sum_i y_{i.}$. It follows that $y_{ij}=\hat\mu+\hat\alpha_i+\hat\epsilon_{ij},\hat\epsilon_{ij}=y_{ij}-\bar y_{i.},\sum_{i=1}^I\hat\alpha_i=0$
One-way F-test
The pooled sample var. $s_p^2=MS_E$ is an unbiased est. of σ^2.
Step 1: Set up hypotheses and determine level of significance H_0: $\mu_1=\cdots=\mu_I$, H_1: Means are not all equal $\mu_u\neq\mu_v$,state α.
Step 2: Select the appropriate test statistic. The test statistic is the F statistic for Anova, F=MSB/MSE.
Step 3: Set up decision rule.
Consider I levels of the main factor A each of sample size J. Degrees of freedom are $df_1=I-1=n_1$ & $df_2=I(J-1)=n_2$. Critical value F_{n_1,n_2} can be found from F distr. table. Decision rule is: Reject H_0 if $F>F_{n_1,n_2}$. Step 4. Compute the test statistic. One-way Anova table.

Source of Variation	Sums of Squares (SS)	Deg. of Freedom (df)	Mean Sqs. (MS)	F
Main factor A	$SS_A=J\sum(\bar y_{i.}-\bar y_{..})^2$	$I-1$	$MS_A=\frac{SS_A}{df_A}$	$F=\frac{MS_A}{MS_E}$
Error (or Residual)	$SS_E=\sum_i\sum_j(y_{ij}-\bar y_{i.})^2$	$I(J-1)$	$MS_E=\frac{SS_E}{df_E}$	
Total	$SS_T=\sum_i\sum_j(y_{ij}-\bar y_{..})^2=SS_A+SS_E$	$IJ-1$		

2. Simultaneous confidence interval

100(1-α)% simultaneous c.i. for a single pair of indep. samples $I_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm t_{df}(\frac{\alpha}{2}) s_p \sqrt{\frac{2}{J}}$.
The multiple comparison problem: the above confidence interval formula is aimed at a single difference, and may produce false discoveries.
Bonferroni method
Bonferroni method is a statistical test repeatedly applied to k independent samples of size n. The overall significance level α is obtained, if each single test is performed at significance level $\alpha_0 = \alpha/k$. Assuming the null hypothesis is true, the number of positive results is $X \sim Bin(k, \alpha_0)$. Thus for small values of α_0, $P(X \geq 1 \mid H_0) = 1-(1-\alpha_0)^k \approx k\alpha_0 = \alpha$. This gives Bonferroni's 100(1-α)% simultaneous confidence interval $B_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm t_{df}(\frac{\alpha}{2k}) s_p \sqrt{\frac{2}{J}}, 1 \leq u < v \leq I$ where $df = I(J-1)$ and $k = \frac{I(I-1)}{2}$. Bonferroni method gives slightly wider intervals compared to the Tukey method.
Tukey method
If I independent samples $(y_{i1}, ..., y_{iJ})$ taken from $N(\mu_i, \sigma^2)$ have the same size J, then $Z_i = \bar{Y}_{i.} - \mu_i \sim N(0, \frac{\sigma^2}{J})$ are independent. Consider the range of differences between Z_i: $R = max\{Z_1, ..., Z_I\} - min\{Z_1, ..., Z_I\}$. The normalised range has a distribution that is free from the parameter σ $\frac{R}{S_p\sqrt{J}} \sim SR(I, df), df = I(J-1)$.
Tukey's 100(1-α)% simultaneous confidence interval is given by $T_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm q_{I,df}(\alpha)\frac{s_p}{\sqrt{J}}$.

3. Kruskal-Wallis test
A nonparametric test, without assuming normality, for no treatment effect H_0: all observations are equal in distribution. Extending the idea of the rank-sum test, consider the pooled sample of size $N = IJ$. Let r_{ij} be the pooled ranks of the sample values y_{ij}, so that $\sum_i \sum_j r_{ij} = 1 + 2 + ... + N = \frac{N(N+1)}{2}$ where the mean rank is $\bar{r}_{..} = \frac{(N+1)}{2}$. Kruskal-Wallis test statistic is given by $W = \frac{12J}{N(N+1)} \sum_{i=1}^{I} (\bar{r}_{i.} - \frac{N+1}{2})^2$.
Reject H_0 for large W using the null distribution table.

4. Two-way layout
$\mu(x_{1i}, x_{2j}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$ where μ is the grand mean,
α_i is the main A-effect, $\sum_{i=1}^{I} \alpha_i = 0$
β_j is the main B-effect, $\sum_{j=1}^{J} \beta_j = 0$
δ_{ij} is the AB-interaction effect, with $\sum_{i=1}^{I} \delta_{ij} = 0$ & $\sum_{j=1}^{J} \delta_{ij} = 0$
Normal theory model
$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}, i = 1, \cdots, I, j = 1, \cdots, J, k = 1, \cdots, K$ where $\epsilon_{ijk} \sim N(0, \sigma^2)$
$\hat{\mu} = \bar{y}_{...}$
$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$
$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$
$\hat{\delta}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$
$\hat{\epsilon}_{ijk} = y_{ijk} - \bar{y}_{ij.}$. Three F-tests-

Source	SS	df	MS	F
Main Effect A		$I-1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_E}$
Main Effect B		$J-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_E}$
Inter. Effect		$(I-1) \times (J-1)$	$\frac{SS_{A*B}}{df_{A*B}}$	$\frac{MS_{A*B}}{MS_E}$
Error		$IJ(K-1)$	$\frac{SS_E}{df_E}$	
Total		$IJK-1$		

Here, $SS_A = JK\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$ & $SS_B = IK\sum_j(\bar{y}_{.j.} - \bar{y}_{...})^2$ & $SS_{A*B} = K\sum_i\sum_j(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$ & $SS_E = \sum(y_i - \bar{y}_{ij})^2$
The mean sums of squares and their expected values $E(MS_A) = \sigma^2 + \frac{JK}{I-1}\sum_i \alpha_i^2$
$E(MS_B) = \sigma^2 + \frac{IK}{J-1}\sum_j \beta_j^2$
$E(MS_{AB}) = \sigma^2 + \frac{K}{(I-1)(J-1)}\sum_i\sum_j \delta_{ij}^2$
$E(MS_E) = \sigma^2$

6. Randomised block design
Blocking is used to remove the effects of the most important nuisance variable. Randomisation is then used to reduce the contaminating effects of the remaining nuisance variables.
Experimental design: randomly assign I treatments within each of J blocks.
Test the null hypothesis of no treatment effect using the two-way layout Anova.
The block effect is anticipated and is not of major interest.
Additive model

Source of Variation	Sums of Squares (SS)	Deg. of Freedom (df)	Mean Sqs. (MS)	F
Main factor A	$SS_A = J\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$I-1$	$MS_A = \frac{SS_A}{df_A}$	F=$\frac{MS_A}{MS_E}$
Main factor B	$SS_B = I\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$	$J-1$	$MS_B = \frac{SS_B}{df_B}$	F=$\frac{MS_B}{MS_E}$
Error (or Residual)	$SS_E = \sum_i\sum_j(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$	$(I-1) \times (J-1)$	$MS_E = \frac{SS_E}{df_E}$	
Total	$SS_T = \sum_i\sum_j(\bar{y}_{ij} - \bar{y}_{..})^2 = SS_A + SS_A + SS_E$	$I(J-1)$		

$E(MS_A) = \sigma^2 + \frac{J}{I-1}\sum_i \alpha_i^2$
$E(MS_B) = \sigma^2 + \frac{I}{J-1}\sum_j \beta_j^2$
$E(MS_E) = \sigma^2$

7. Friedman test
Here we introduce another nonparametric test, which does not require that ϵij are normally distributed, for testing H_0: no treatment effect. The Friedman test is based on within block ranking. Let ranks within j-th block be: $(r_{ij}, \cdots, r_{Ij})$=ranks of $(r_{ij}, \cdots, r_{Ij})$ so that $r_{1j} + \cdots + r_{Ij} = 1 + 2 + ... + I = \frac{I(I+1)}{2}$ where $\frac{1}{J}(r_{1j} + \cdots + r_{Ij}) = \frac{I+1}{2}$ and $\bar{r}_{..} = \frac{(I+1)}{2}$. Friedman test statistic $Q = \frac{12J}{I(I+1)}\sum_{i=1}^{I}(\bar{r}_{i.} - \frac{I+1}{2})^2$ has an approximate null distribution $Q \sim \chi^2_{I-1}$. Reject H_0 for large W using the null distribution table.

IX Categorical data analysis

1. Chi-squared test of homogeneity
Consider a table of I × J observed counts obtd. from J indep. samples taken from J pop. distr.:

	Pop. 1	Pop. 2 $\cdots$	Pop. J	Total	
Category 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
Category 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Category I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
Sample sizes	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	$n_{I.}$

This model is described by J multinomial distr. $(N_{1j}, \cdots, N_{1j}) \sim Mn(n_{.j}; \pi_{1|j}, \cdots, \pi_{I|j}), j = 1, \cdots, J$ Under the hypothesis of homogeneity $H_0 : \pi_{i|j} = \pi_i, \forall(i,j)$, the m.l.e of π_i are the pooled sample proportions $\hat{\pi}_i = \frac{n_{i.}}{n_{..}}, i = 1, \cdots, I$. Using m.l.e, we compute the expected cell counts $E_{ij} = n_{.j}\hat{\pi}_i = \frac{n_{i.}n_{.j}}{n_{..}}, i = 1, \cdots, I$ & the chi-squared test statistic becomes $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{i.}n_{i.}/n_{..})^2}{n_{i.}n_{i.}/n_{..}}$ We reject H_0 for large values of χ^2 & $df = (I-1)(J-1)$.

2. Chi-squared test of independence
Data: a single cross-classifying sample is summarised in terms of the observed counts, whose joint distribution is multinomial$(N_{1j}, \cdots, N_{1j}) \sim Mn(n_{.j}; \pi_{1|j}, \cdots, \pi_{I|j}), j = 1, \cdots, J$ Under the hypothesis of homogeneity $H_0 : \pi_{i|j} = \pi_i, \forall(i,j)$, the m.l.e of π_i are the pooled sample proportions $\hat{\pi}_i = \frac{n_{i.}}{n_{..}}, i = 1, \cdots, I$. Using m.l.e, we compute the expected cell counts $E_{ij} = n_{.j}\hat{\pi}_i = \frac{n_{i.}n_{.j}}{n_{..}}, i = 1, \cdots, I$ & the chi-squared test statistic becomes $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{i.}n_{i.}/n_{..})^2}{n_{i.}n_{i.}/n_{..}}$
We reject H_0 for large values of χ^2 & $df = (I-1)(J-1)$.

3. Matched-pairs designs
McNemar's test
Consider data obtd. by matched-pairs design for pop. distr.

	unaffected X	unaffected $\bar{X}$	Total
affected X	p_{11}	p_{12}	$p_{1.}$
affected $\bar{X}$	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

Null hypothesis $H_0 : p_{12} = p_{21} = p$ This yields the McNemar's test statistic: $\chi^2 = \frac{(m_{12} - m_{21})^2}{m_{12} + m_{21}}$ whose approx. null distr. is χ^2_1

4. Odds ratios
Conditional odds for A given B are defined as $odds(A|B) = \frac{P(A|B)}{P(\bar{A}|B)} = \frac{P(AB)}{P(\bar{A}B)}$.
Odds ratio for a pair of events defined by $\Delta_{AB} = \frac{odds(A|B)}{odds(A|\bar{B})} = \frac{P(AB)P(\bar{A}\bar{B})}{P(\bar{A}B)P(A\bar{B})}$. The odds ratio is a measure of dependence between a pair of random events having the following properties
if $\Delta_{AB} = 1$, then events A and B are independent,
if $\Delta_{AB} > 1$, then $P(A|B) > P(A|\bar{B})$ & so B inc. prob. of A,
if $\Delta_{AB} < 1$, then $P(A|B) < P(A|\bar{B})$ & so B dec. prob. of A.
Odds ratios for case-control studies
• Conditional probabilities-

	X	$\bar{X}$	Total
D	$P(X\|D)$	$P(\bar{X}\|D)$	1
$\bar{D}$	$P(X\|\bar{D})$	$P(\bar{X}\|\bar{D})$	1

Corresponding odds ratio $\Delta_{DX} = \frac{P(X|D)P(\bar{X}|\bar{D})}{P(\bar{X}|D)P(X|\bar{D})}$
• Observed counts-

	X	$\bar{X}$	Total
D	n_{11}	n_{12}	$n_{1.}$
$\bar{D}$	n_{21}	n_{22}	$n_{2.}$

Corresponding odds ratio $\hat{\Delta}_{DX} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

X Multiple regression

• Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables. For instance, when we predict rent based on square feet alone that is simple linear regression. When we predict rent based on square feet and age of the building that is an example of multiple linear regression.

1. Simple linear regression model
A simple linear regression model connects two random vars (X,Y): X is called predictor variable and Y is called response by a linear relation $Y = \beta_0 + \beta_1 X + \epsilon$, $\epsilon \sim N(0, \sigma^2)$ where ϵ is the noise. The key assumption of the model requires that has a normal distribution $N(0, \sigma^2)$ indep. of X. This assumption is called homoscedasticity, meaning that the noise size σ is the same for all possible levels of the predictor var.
The fitted regression line is
$y = b_0 + b_1 x = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$ where sample correlation coefficient is $r = \frac{s_{xy}}{s_x s_y}$, $s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$.

2. Residuals
Then the size of noise (estimated σ^2) is $s^2 = \frac{n-1}{n-2}s_y^2(1 - r^2)$.
Decomposition: $y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i$ implies $SS_T = SS_R + SS_E$
$SS_T = \sum_i(y_i - \bar{y})^2 = (n-1)s_y^2$
$SS_R = \sum_i(\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2$
$SS_E = S(b_0, b_1) = \sum_i(y_i - \hat{y}_i)^2 = (n-1)s_y^2(1 - r^2)$.
Combining them, $r^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$. Thus the squared sample correlation coefficient r^2 is called the coefficient of determination. Coefficient of determination r^2 is the proportion of variation in the response variable explained by the variation of the predictor. r^2 is independent of choice of the explanatory and the response variables.
To test the normality assumption, use the normal distribution plot for the standardised residuals $\tilde{e}_i = \frac{\hat{e}_i}{s_i}, i = 1, \cdots, n$ where $s_i = s\sqrt{1 - \frac{\sum_k(x_k - x_i)^2}{n(n-1)s_x^2}}$.
For simple linear regression model, scatter plot of the standardised residuals versus x_i should look as a horizontal blur. Non-linearity problem is fixed by log-log transformation of the data.

3. Confidence intervals and hypothesis testing
$i = 0, 1, B_i \sim N(\beta_i, \sigma_i^2)$, $s_{b_i}^2 = \frac{s^2\sum x_i^2}{n(n-1)s_x^2}$, $\frac{B_i - \beta_i}{S_{B_i}} \sim t_{n-2}$.
Exact 100(1-α)% c.i. $I_{\beta_i} = bi \pm t_{n-2}(\frac{\alpha}{2})s_{b_i}$.
Test statistic $t = \frac{b_i - \beta i}{s_{b_i}}$ that has the exact null distr. $T \sim t_{n-2}$.

4. Intervals for individual observations
• In cases where we assume a model with an explanatory variable X and a response variable Y, the prediction interval is an interval in which a single sample of Y falls with a certain probability for a given X. In this context, a confidence interval corresponds to an interval which the mean of Y for a given X falls with a certain probability. Thus the prediction interval is always wider than the confidence interval for a given signficance level.
Exact confidence interval $I_\mu = b_0 + b_1 x \pm t_{n-2}(\frac{\alpha}{2})s\sqrt{\frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2}$
Exact prediction interval $I_\mu = b_0 + b_1 x \pm t_{n-2}(\frac{\alpha}{2})s\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2}$
Prediction interval has wider limits since it contains the uncertainty due the noise factors: $Var(Y - \hat{\mu}) = Var(\mu + \epsilon - \hat{\mu}) = \sigma^2 + Var(\hat{\mu}) = \sigma^2(1 + \frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2)$

5. Linear regression and ANOVA
6. Multiple linear regression
Multiple linear regression model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1} + \epsilon$, $\epsilon \sim N(0, \sigma^2)$
Corr. data set consists of n indep. vectors with $n > p$ is
$y_1 = \beta_0 + \beta_1 x_{1,1} + \cdots + \beta_{p-1}X_{1,p-1} + e_1$
$y_n = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_{p-1}X_{n,p-1} + e_n$
Design matrix
$1\ x_{1,1} \cdots x_{1,p-1}$
$\cdots$
$1\ x_{n,1} \cdots x_{n,p-1}$
An unbiased estimate of σ^2 is $s^2 = \frac{SS_E}{n-p}$ where $SS_E = ||\hat{e}||^2 = ||y - \hat{y}||^2$.
To check the underlying normality assumption inspect the normal probability plot for the standardised residuals $\frac{\hat{e}_i}{s\sqrt{1-p_{ii}}}$ where p_{ii} are the diagonal elements of P.
Coefficient of multiple determination $R^2 = 1 - \frac{SS_E}{SS_T}, SS_T = (n-1)s_y^2$.
Adjusted coefficient of multiple determination $R_a^2 = 1 - \frac{n-1}{n-p}\frac{SS_E}{SS_T} = 1 - \frac{s^2}{s_y^2}$. The adjustment factor $\frac{n-1}{n-p}$ gets larger for the larger values of predictors p.

Basic probability theory

• Cumulative distribution function $F(x) = P(X \leq x) = \sum_{y \leq x} f(y) or = \int_{y \leq x} f(y)dy$
• Expected value (mean or average) of X $\mu = E(X) = \sum_x f(x) or = \int_x f(x)dx$
• $Var(cX) = c^2 Var(X)$
• $Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2$
• Standard normal distribution $Z \sim N(0,1)$ has the density fn. and distribution fn. $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} = \frac{1}{\sqrt{2\pi}}e^{-(\frac{X-\mu}{\sigma})^2/2}$, $\phi(z) = \int_{-\infty}^{z}\phi(x)dx$
• Normal distribution $X \sim N(\mu, \sigma^2)$: $\frac{X-\mu}{\sigma} \sim N(0,1)$, $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, $E(X) = \mu$, $Var(X) = \sigma^2$
• Discrete uniform distribution $X \sim U(N)$: $f(k) = \frac{1}{N}, 1 \leq k \leq N, E(X) = \frac{N+1}{2}, Var(X) = \frac{N^2-1}{12}$
• Continuous uniform distribution $X \sim U(a,b)$: $f(k) = \frac{1}{b-a}, a < x < b, E(X) = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}$
• Binomial distribution $X \sim Bin(n,p)$: $f(k) = \binom{n}{k}p^k(1-p)^{n-k}, 0 \leq k \leq n, E(X) = np, Var(X) = np(1-p)$
• Bernoulli distribution $Ber(p) = Bin(1,p)$
• Geometric distribution $X \sim Geom(p)$: $f(k) = pq^{k-1}, k \geq 1, E(X) = \frac{1}{p}, Var(X) = \frac{1-p}{p^2}$
• Exponential distribution $X \sim Exp(\lambda)$: $f(x) = \lambda e^{-\lambda x}, x > 0, E(X) = \sigma_X = \frac{1}{\lambda}$
• Poisson distr. $X \sim Pois(\lambda)$: $f(k) = \frac{\lambda^k}{k!}e^{-\lambda}, k \leq 0, E(X) = Var(X) = \lambda$
The two-sample t-test assumes that two independent samples $(X_1, ..., X_n)$ and $(Y_1, \cdots, Y_m)$ are taken from two normal distributions with equal variance. To test this normality assumption one may use a normal probability plot for n+m residuals $X_1 - \bar{X}, \cdots, X_n - \bar{X}, Y_1 - \bar{Y}, \cdots, Y_m - \bar{Y}$.
Without taking account of multiple comparisons the CI is much narrower producing an excess of false positive results.
Multiple regression-The normality assumption can be justified in the case when the noise value is the sum of many independent and relatively small factors. Equal variance is realistic if the external factors are more or less the same across the three different experiments.