Statistics

Weber

July 5, 2008

1 Paramater estimation

1.1 What is statistics?

The lecturer defines stasistics as a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.

It has two branches; estimation and hypothesis testing.

Example 1.1

Say we want to estimate p =proportion of cambridge students who've not bathed or showered within the last 24 hours.

Numbers alone are just numbers; they become data (a plural noun) when we know what they represent.

1.2 Random Variables with values in $\mathbb{R}^n$ or $\mathbb{Z}^n$

$X = (X_1, \ldots, X_n)$ where X_i takes values in $\mathbb{R}$ or $\mathbb{Z}$; data is of the form $x = (x_1, \ldots, x_n)$.

Recall that an RV is a function $X : \Omega \to \mathbb{Z}$; Ω is the sample space, e.g. when tossing two coins $\Omega = \{HH, HT, TH, TT\}$ and $X(\omega) :=$ no. of heads when the outcome is ω.

The distribution function $F_X(x) := P(X \leq x)$; for X discrete this is $\sum_{\omega : X(\omega) \leq x} P(\omega)$, for X continuous it is $\int_0^x f(u)du$ where f is the probability density function. $EX = \sum_\omega X(\omega)P(\omega)$ or $\int_{-\infty}^{\infty} f(u)udu$; $Eh(x) = \int h(u)f(u)du$. $\mathrm{Var}(x) = E(X - EX)^2 = EX^2 - (EX)^2$. We often write $EX = \mu, \mathrm{Var}(x) = \sigma^2$.

1.3 Some important RVs

a) $X \sim B(n, p)$, the binomial distn; $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, 0 \leq k \leq n$, $EX = np$, $\mathrm{Var}(X) = np(1-p)$

b) $X \sim P(\lambda)$, the Poisson distn; $P(X = k) = \lambda^k \frac{e^{-\lambda}}{k!}, k = 0, 1, \ldots,$ $EX = \lambda$, $\mathrm{Var}(X) = \lambda$

Also important are the Normal, Standard Normal and Uniform distributions for continuous variables; see the printed notes for this course.

1.6 Notion of a Statistic

Say we have $x_1, \ldots, x_n$ data drawn as IID samples from some dist, e.g. $N(\mu, \sigma^2)$. A statistic is a function $T(x)$, such as $\max\{x_i\}, \frac{x_1+x_2}{x_3}, \log x_3, 2007+10\min\{x_i\}$. Say μ is unknown; $\frac{1}{n}(x_1 + \cdots + x_n)$ s a good statistic to use to estimate it, but why is this a better estimate than any of the other statistics?

1.7 Unbiased Estimators

An estimator of an unknown paramater θ [i.e. a statistic used to estimate θ] is unbiased if $ET(X) = \theta$, e.g. if $X_1, \ldots, X_n$ are i.i.d. as $B(1, p)$ with p unknown we define $\hat{p}(X) = \frac{1}{n}\sum X_i$ and then $E\hat{p}(X) = E(\frac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n}(EX_1 + \cdots + EX_n) = \frac{np}{n} = p$ so $\hat{p}$ is unbiased.

These are **not** generally unique; $\tilde{p}(X) = \frac{1}{3}X_1 + \frac{2}{3}X_2$ is also unbiased, but we intuitively "know" this is a "worse" estimator. One reason for tihs is than $\mathrm{Var}(\hat{p}) < \mathrm{Var}(\tilde{p})$.

Some more important RVs are the Geometric, Exponential and Gamma distributions; new to some readers will be the Beta distribution.

Before proceeding any further, the reader should ensure they are familiar with the Weak and Strong Laws of Large Numbers, and the Central Limit Theorem, from last year's Probability course.

2 Maximum Likelihood estimation

2.1 Estimation

Say $X_1, \ldots, X_n$ i.i.d. RVs, $x_1, \ldots, x_n$ data. The RVs are $N(\mu, \sigma^2), B(n, p)$ or $P(\lambda)$ w/ paramaters to be estimated.

Define likelihood$(\theta) = f(x \mid \theta)$ for fixed $x = (x_1, \ldots, x_n)$, where $f(x_i \mid \theta) =$ is the pdf at x_i [of θ ?], $f(x \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta) = \mathrm{like}(\theta)$. The maximum likelihood estimate at θ is the value at θ maximising like(θ), say $\tilde{\theta}(x)$. It is often convenient to maximise $\log \mathrm{like}(\theta)$, called log-likelihood.

Example 2.1

a) How many colours do Smarties come in? Suppose k colours, all equally likely. Suppose we take 3 smarties and these are red, green, red. Let $x = \{$2nd different from 1st and 3rd same as 1st$\}$. like$(k) = P(x \mid k) = \frac{k-1}{k} \times \frac{1}{k} = \frac{k-1}{k^2}$. For $k = 2, 3, 4, \ldots,$ like$(k) = \frac{1}{4}, \frac{2}{9}, \frac{3}{16}, \ldots;$ $k(x) = 2$ maximizes. Suppose the 4th is orange; like$(k) = \frac{k-1}{k^2}\frac{k-2}{k} = \frac{(k-1)(k-2)}{k^3} = \frac{2}{27}, \frac{3}{32}, \frac{12}{125}, \frac{5}{54}, \ldots;$ the max here is at $k = 5$.

b) $X \sim B(n, p)$; $\log p(x \mid n, p) = \log \binom{n}{x} p^x (1-p)^{n-x}$; n is known, p unknown, so this $= \cdots + x \log p + (n - x)\log(1 - p)$. $\frac{\partial}{\partial p} = 0 \Rightarrow \frac{x}{p} - \frac{n-x}{1-p} = 0 \Rightarrow \hat{p}(x) = \frac{x}{n} \Rightarrow \hat{p}(X) = \frac{X}{n}$. This is unbiased; $E(\hat{p}(X)) = \frac{EX}{n} = \frac{np}{n} = p$

c) $X \sim B(n, p)$, p known, n unknown. $P(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$, to be maximised wrt n for $n \in \{x, x+1, x+2, \ldots\}$. $\frac{P(x|n+1,p}{P(x|n,p)} = \frac{\binom{n+1}{x} p^x (1-p)^{n+1-x}}{\binom{n}{x} p^x (1-p)^{n-x}} =$

$\frac{(1-p)(n+1)}{n+1-x}$; if we graph this we see $n+1-x \leq 1 \Leftrightarrow n+1 \geq \frac{x}{p}$. $A(x) = \left[\frac{x}{p}\right]$; if $\frac{x}{p} \in \mathbb{Z}$ then $\frac{x}{p}$ and $\frac{x}{p-1}$ are both MLEs

d) $X_1, \ldots, X_n \sim$ geometric(p); $\log p(x_1, \ldots, x_n \mid p) = \log \prod_{i=1}^n (x_i \mid p) = \log \prod_{i=1}^n (1-p)^{x_i - 1} p = (\sum x_i - n)\log(1-p) + n \log p$; $\frac{\partial}{\partial p}(\ldots) = 0 \Rightarrow \frac{-(\sum x_i - n)}{1-p} + \frac{n}{p} = 0$; MLE $\hat{p} = \frac{1}{\bar{x}}$ ($\hat{x} = \frac{\sum x_i}{n}$; $E(\hat{p}(x)) = E\left(\frac{1}{\bar{x}}\right) \neq \frac{E1}{E\bar{x}}$; for the case $n = 1$ $E\hat{p}(x) = E\frac{1}{x_1} = \sum_1^\infty \frac{1}{j}(1-p)^{j-1}p = -\frac{p}{1-p}\log p > p$, so this estimator is biased

2.2 Sufficient Statistics

$\bar{x} = \frac{\sum x_i}{n}$. $T(x)$ is said to be sufficient for θ if $p_\theta(x \in . \mid T(X) = t)$ doesn't depend on θ.

Thm 2.2

The statistic T is sufficient for θ iff $f(x \mid \theta) = g(T(x), \theta)(x)$; this is called the factorisation criteria. Suppose the sample space is discrete and $f(x \mid \theta) = p_\theta(X = x)$ has the factorisation criteria, then $p_\theta(X = x \mid T(X) = t) = \frac{p_\theta(X=x)}{p_\theta(T(X)=t)} = \frac{g(T(x),\theta)h(x)}{\sum_{x:T(x)=t} g(T(x),\theta)h(x)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}$ which does not depend on θ. $p(x \mid \theta) = p_\theta(X = x) = p_\theta(T(X) = t)p_\theta(X = x \mid T(X) = t) = g(T(X), \theta)h(x)$.

Example 2.3

a) $X_1, \ldots, X_n \sim P(\lambda)$; $f(x \mid \lambda) = \prod_1^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda^{\sum x_i} e^{-n\lambda} \prod \frac{1}{x_i!} = g(\sum x_i, \lambda)h(x)$, so $\sum x_i$ is sufficient for λ. Note MLE of λ must depend on $\sum x_i = T(X)$; $\tilde{\lambda}\,\text{MLE}(X) = \frac{T(X)}{n} = \bar{X}$

b) $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$; $\theta = (\mu, \sigma^2)$ to be estimated. $f(x \mid \mu\sigma^2) = \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\left[\sum (x_i - \bar{x})^2 + n(\bar{x}-\mu)^2\right]} = s(\sum (x_i - \bar{x})^2, \bar{x}, \mu, \sigma^2)h(x)$; $T(X) = (\bar{x}, \sum (x_i - \bar{x})^2)$, $\hat{\mu} = \bar{x}$, $\hat{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{n}$

c) $X_1, \ldots, X_n \sim U[0, \theta]$, $f(x \mid \theta) = \prod_{i=1}^n I[0 \leq x_i \leq \theta]\frac{1}{\theta} = \frac{1}{\theta^n} I[\max_i x_i \leq \theta]$. We want this to $= g(T(x), \theta)h(x)$ so $T(x) = \max x_i$ is sufficient for θ; the MLE is $\hat{\theta}(x) = \max x_i$. $E\hat{\theta}(X) = E \max x_i$; $P(\max X_i \leq t) = F(t) = P(X_1 \leq t, \ldots, X_n \leq t) = P(X_1 \leq t) \ldots P(X_n \leq t) = \left(\frac{t}{\theta}\right)^n$, $f(t) = F'(t) = \frac{nt^{n-i}}{\theta^n}$, so $E \max X_i = \int_0^\theta t f(t) dt = \int_0^\theta \frac{tnt^{n-1}}{\theta^n} dt = \frac{n}{n+1}\theta \neq \theta$ so the MLE is biased, however notice it is asymtotically unbiased, i.e. $E\hat{\theta} \to \theta$ as $n \to \infty$

3 The Rao-Blackwell Theorem

3.1 Mean square error

A good estimator should make small $E((\hat{\theta}(X) - \theta)^2)$ (which is of course the variance of $\hat{\theta}$ for the case when $\hat{\theta}$ is unbiased), the mean square error.

Example 3.1

$X_1, \ldots, X_n \sim B(1,p)$, p to be estimated. $\hat{p} = \bar{X} = \frac{X_1+\cdots+X_n}{n}, \tilde{p} = \frac{X_1+2X_2}{3}$; these are both unbiased so we compare variance; $\mathrm{Var}(\hat{p}) = \frac{1}{n^2}\sum \mathrm{Var}(x_i) = \frac{p(1-p)}{n}, \mathrm{Var}(\tilde{p}) = \frac{1}{9}(\mathrm{Var}(x_1) + 4\,\mathrm{Var}(x_2) = \frac{5}{9}p(1-p)$. There are also unbiased estimators for which the variance decreases with increasing n but not as rapidly as for $\hat{p}$, e.g. $p^\star = \frac{X_1+2X_2+\cdots+nX_n}{\frac{n(n+1)}{2}}$ has $Ep^\star = p$ and we can find $\mathrm{Var}(p^\star) = \frac{2(2n+1)}{3n(n+1)}p(1-p)$ and $\frac{\mathrm{Var}(\hat{p})}{\mathrm{Var}(p^\star)} \to \frac{3}{4}$ as $n \to \infty$.

Example 3.2

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, both paramaters unknown. $\log(f(x)\mu\sigma^2) = \log\prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_1^n (x_i-\mu)^2$; for unbiased estimators the partial derivatives of this wrt μ, σ^2 must be both 0 i.e. $\frac{1}{\hat{\sigma}^2}\sum(X_i-\hat{\mu}) = 0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^2}\sum(X_i-\hat{\mu})^2$. So $\hat{\mu} = \frac{\sum X_i}{n} = \bar{X}, \hat{\sigma}^2 = \frac{1}{n}\sum(X_i-\hat{\mu})^2 = \frac{1}{n}\sum(X_i-\bar{X})^2$; $E\hat{\mu} = \mu$ so this is unbiased but $E\sum_1^n(X_i-\bar{X}) = (n-1)\sigma^2$ so $E\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$ so this is biased.

We might prefer to have an unbiased estimator; we see we can just multiply by a constant, $\tilde{\sigma}^2 = \frac{1}{n-1}\sum(X_i-\bar{X})^2$ is unbiased. However, neither the MLE nor this unbiased estimator minimizes the mean square error; for an estimator of the form $\lambda\sum(X_i-\bar{X})^2 := S_{XX}$, the MSE is $E((\lambda S_{XX}-\sigma^2)^2) = \lambda^2 ES_{XX}^2 - 2\lambda\sigma^2 ES_{XX} + \sigma^4$; we already know $ES_{XX} = (n-1)\sigma^2$ and will later find ES_{XX}^2, which gives that this is $\lambda^2(2(n-1)\sigma^4 + (n-1)^2\sigma^4) - 2\lambda\sigma^2(n-1)\sigma^2 + \sigma^4$ minimised by $\lambda = \frac{1}{n+1}$.

3.2 Rao-Blackwell Theorem

Thm 3.3

Let $\hat{\theta}$ be an estimator at θ w/ $E\hat{\theta}^2 < \infty \forall\theta$. Suppose T is a sufficient statistic for θ, and let $\theta^\star = E(\hat{\theta} \mid T)$, then $E(\theta^\star-\theta)^2 \le E(\hat{\theta}-\theta)^2$, w/ equality only if $\hat{\theta}$ is a function of T.

The proof is just a few lines, but conceptually somewhat difficult: $E(\theta^\star - \theta)^2 = E(E(\hat{\theta} \mid T) - \theta)^2 = E(E(\hat{\theta}-\theta \mid T))^2 \le E(E((\hat{\theta}-\theta)^2 \mid T)) = E(\hat{\theta}-\theta)^2$ ($\forall w, (Ew)^2 \le Ew^2$ since $\mathrm{Var}(w) = Ew^2 - (Ew)^2 \ge 0$, with equality only when $\mathrm{Var}(w) = 0$ i.e. w constant) with equality only when $\hat{\theta} - \theta \mid T$ is a constant i.e. $\hat{\theta}$ is a function of T.

If $\hat{\theta}$ is unbiased, $E(\theta^\star) = E(E(\hat{\theta} \mid T)) = E\hat{\theta} = \theta$ so $\theta^\star$ is also unbiased.

Examples 3.4

a) $X_1, \ldots, X_n \sim P(\lambda)$; a sufficient statistic for λ is $\sum X_i$. We start with a trivial estimator $\tilde{\lambda} = X_1$. $\lambda^\star = E(\tilde{\lambda} \mid T) = E(X_1 \mid \sum_1^n X_i = t)$. Now $E(\sum_1^n X_i \mid \sum X_i = t) = t = \sum_{j=1}^n E(X_j \mid \sum_i X_i = t)$, so this is $\frac{t}{n}$

b) $X_1, \ldots, X_n \sim P(\lambda), \theta = e^-\lambda$ to be estimated. $\theta = P(X_1 = 0) \therefore \hat{\theta} = I[X_1 = 0]$ is unbiased. $\theta^\star(t) = E(\hat{\theta} \mid T = t) = P(X_1 = 0 \mid \sum_1^n X_i = t) = \frac{P(X_1=0 \text{ and } \sum_1^n X_i = t)}{P(\sum_1^n X_i = t)} = \frac{\frac{e^{-\lambda}(\lambda(n-1))^t e^{-\lambda(n-1)}}{t!}}{\frac{(\lambda n)^t e^{-\lambda n}}{t!}} = \left(\frac{n-1}{n}\right)^t$, so $\theta^\star(X) = \left(\frac{n-1}{n}\right)^{\sum X_i}$

c) $X_1, \ldots, X_n \sim U[0,\theta]$; $EX_1 = \frac{\theta}{2}$ so $2X_1$ is unbiased; $\theta^\star = E(\hat{\theta} \mid T = t)$; $T = \max x_i$ is sufficient for θ so this is $E(2X_1 \mid \max X_i = t) = \frac{1}{n}2t + \frac{n-1}{n}2\frac{t}{2} = \frac{n+1}{n}t$

3.3 Consistency and asymtotically efficient

MLEs are always asymtotically unbiased, though we will not proove this in this course: $E(\hat{\theta}_{\text{MLE}}) \to \theta$ as $n \to \infty$. In fact, we have a stronger property, called consistency, which is that $P(|\hat{\theta}_{\text{MLE}} - \theta| > t) \to 0$ as $n \to \infty$. We also have that $\lim_{n\to\infty} \frac{\text{Var}(\hat{\theta}_{\text{MLE}})}{\frac{1}{nI(\theta)}} \to 1$ as $n \to \infty$, with the denominator being the Crawer-Rao lower bound on the variance of an estimator; the MLE is asymtotically efficent.

4 Confidence intervals

4.1 Interval estimates

Say $X_1, \ldots, X_n \sim N(\theta, 1)$ and we have an unbiased estimator $\hat{\theta}$. Even if $E(\hat{\theta}(X) - \theta)^2$ is small, we will often have $\hat{\theta} \neq \theta$. We can instead consider the probability that $\theta \in$ an interval estimate $[a(x), b(x)]$; $[-\infty, \infty]$ is correct w/ prob 1. We define an interval estimator $[a(X), b(X)]$; if $P([a(X), b(X)] \ni \theta) = \gamma$ this defines a $\gamma \times 100\%$ confidence interval for θ.

Example 4.1

a) $X_1, \ldots, X_n \sim N(\mu, \sigma^2), \mu$ unknown. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n}$ $\therefore \bar{X} - \mu \sim N(0, \frac{\sigma^2}{n} \Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$. We want $P(\xi \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq \eta) = 0.5 = P(\bar{X} - \frac{\eta\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{\xi\sigma}{\sqrt{n}}$; $\bar{X} - \frac{\eta\sigma}{\sqrt{n}} = a(X), \bar{X} + \frac{\xi\sigma}{\sqrt{n}} = b(X)$. We want to minimize $b(X) - a(X)$, which we do by choosing a symetrical interval; for $W \sim N(0,1), P(-1.96 \leq W \leq 1.96) = 0.95, P(-2.58 \leq W \leq 2.58) = 0.99$

b) $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with both unknown. $\frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\frac{S_{XX}}{n-1}}} \sim t_{n-1}$ where the RHS is the students' t-distribution and $S_{XX} = \sum(X_i - \bar{X})^2$. $\gamma = 0.95 = P(\xi \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\frac{S_{XX}}{n-1}}} \leq \eta) = P(\bar{X} - \eta\sqrt{\frac{S_{XX}}{n(n-1)}} \leq \mu \leq \bar{X} + \xi\sqrt{\frac{S_{XX}}{n(n-1)}}$; notice $t_\infty = N(0,1)$

4.2 Opinion Polls

Let $p =$ probability someone supports Labour; $X_i \sim B(1,p)$ are 1 if a person supports Labour, 0 otherwise. $\hat{p} = \frac{1}{n}(X_1 + \cdots + X_n)$. $X_1 + \cdots + X_n \sim B(n,p) \approx\sim N(np, np(1-p))$, so $\bar{X} \approx\sim N(p, \frac{p(1-p)}{n}$. $E\bar{X} = p, \text{Var}\,\bar{X} = \frac{p(1-p)}{n} \leq \frac{1}{n}$ (p is unknown, but the variance is maximised by $p = \frac{1}{2}$), so $\frac{(\bar{X}-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$. $P(\hat{p} - 0.03 \leq p \leq \hat{p} + 0.03) = P(\frac{-0.03}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0.03}{\sqrt{\frac{p(1-p)}{n}}}$. $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$ so this is $\Phi(0.03\sqrt{\frac{n}{p(1-p)}} - \Phi(-0.03\sqrt{\frac{n}{p(1-p)}} \geq \Phi(0.03\sqrt{4n} - \Phi(-0.03\sqrt{4n}$, which is ≥ 0.95 if $0.03\sqrt{4n} \geq 1.96 \Leftrightarrow n \geq 1068$; for real opinion polls $n = 1100$ is used, regardless of the population size.

Example 4.2

Of 1000 Americans, 59% believe the world will end, and of those, 33% believe it will within a decade, therefore this is 19.5% of the population.

Rule of 39

[I found this section incomprehensible]

Opinion Polls

$\mathrm{Var}(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n}$, where N is the total population.

Rk

μ, p location
σ scale

Example 4.3

$X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$. $f(x_1, \ldots, x_n \mid \theta) = \prod_1^n \hat{\theta} e^{-\theta x_i} = \theta^n e^{-\theta \sum x_i}$, so $T(X) = \sum X_i$ is sufficient for θ. $\sum X_i \sim \Gamma(n, \theta)$, $f_T(t) = \frac{\theta^n t^{n-1} e^{-\theta t}}{(n-1)!}, t \geq 0$. $S = 2\theta T \sim \Gamma(n, \frac{1}{2})$. $P(S \leq s) = P(2\theta T \leq s)$. $f_S(s) = f_T(\frac{s}{2\theta})\frac{1}{2\theta} = \frac{\theta^n (\frac{S}{2\theta})^{n-1} e^{-\theta \frac{S}{2\theta}}}{(n-1)!} = \frac{(\frac{S}{2})^{n-1} e^{-\frac{S}{2}}}{(n-1)!}$. $P(\xi \leq 2T\theta \leq 2\eta) = P(\frac{2T}{\eta} \leq \frac{1}{\theta} \leq \frac{2T}{\xi}) = F_{2n}(\xi) - F_{2n}(\eta)$ where F_{2n} is the cdf of a χ^2_{2n} random variable.

4.3 Shortcomings of CI

$X_1, X_2 \sim U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. $P(\min X_i \leq \theta \leq \max X_i) = P(X_2 \leq \theta \leq X_1) + P(X_1 \leq \theta \leq X_2) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}$, so $[\min x_i, \max x_i]$ is 50% CI. But if e.g. $X = (7.4, 8.0)$ then $\theta \leq 7.4 + \frac{1}{2} = 7.9$ and similarly $\theta \geq 7.5$ so $[7.4, 8.0]$ is a 100% CI, not 50%.

5 Bayesian Estimation

5.1 Prior and Posterior Distributions

In Bayesian statistics we take the view that a probability represents our level of belief in a given proposition, and requires us to incorporate our prior beliefs, in the form of a prior distribution for θ. We then combine data with this to get posterior beliefs - the beliefs we hold after seeing the data.

$P(\theta \mid \text{action}) = P(\theta \mid x_1 \ldots x_n) = \frac{f(x_1 \ldots x_n \mid \theta) p(\theta)}{\int f(x_1 \ldots x_n \mid \theta) p(\theta) d\theta} \propto f(x_1 \ldots x_n \mid \theta) p(\theta)$.

Example 5.1

Take our prior distribution for the number of colours of smarties to be 5,6,7,8 with respective probabilities $\frac{1}{10}, \frac{3}{10}, \frac{3}{10}, \frac{3}{10}$. If our data is $x =$red, green, red, let $\theta = k$ be the no. of colours. $f(x \mid k) = \frac{k-1}{k^2}$ so we have:

k	$f(x \mid k)$	$f(x \mid k)p(k)$	Posterior $p(k \mid x)$
5	0.160	0.016	0.13
6	0.139	0.042	0.33
7	0.122	0.037	0.29
8	0.109	0.033	0.26
	total	0.127	

Similarly, if our data is red, green, orange, we have:

k	$f(x \mid k)$	$f(x \mid k)p(k)$	Posterior $p(k \mid x)$
5	0.096	0.010	0.11
6	0.093	0.028	0.31
7	0.087	0.026	0.30
8	0.082	0.025	0.28
	total	0.088	

5.2 Conditional PDFs

Discrete case

The key idea here is that $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ (or 0 if $P(B) = 0$); write $f_{XY}(x,y) = P(X = x, Y = y), f_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X=x,Y=y}{P(Y=y)} = \frac{f_{XY}(x,y)}{f_Y(y)}$, or 0 if $f_Y(y) = 0$.

Example 5.2

$X \sim \text{Poisson}(\lambda), R \sim \text{Poisson}(\mu)$ independent, $Y = X + R \sim \text{Poisson}(\lambda + \mu)$. $f_{X|Y}(x,y) = \frac{\frac{\lambda^x e^{-\lambda}}{x!}\frac{\mu^{y-x}e^{-\mu}}{(y-x)!}}{\frac{(\lambda+\mu)^y}{y!}e^{-(\lambda+\mu)}} = \binom{y}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(1-\frac{\lambda}{\lambda+\mu}\right)^{y-x}$, i.e. $X \mid Y$ is distributed as $B(y, \frac{\lambda}{\lambda+\mu})$.

Continuous case

$Z = (X,Y), f_Z(x,y) = f_{X,Y}(x,y), f_Y(t) = \int f_{X,Y}(x,y)dx, f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, or 0 if $f_Y(y) = 0$.

Example 5.3

a) $\theta =$ prob. of heads on a biased coin. Let $x_1, \ldots, x_n$ each be 1 for heads, 0 for tails, and $\sum x_i = t$. Let our prior distribution be $p(\theta) = 1 \forall 0 \le \theta \le 1$; $P(\theta \mid x_1 \ldots x_n) \propto \theta^t(1-\theta)^{n-t}$ (only the parts in terms of θ are relevant [lulz]. $p(\theta \mid x_1 \ldots x_n) = \frac{\theta^t(1-\theta)^{n-t}}{\int_0^1 \theta^t(1-\theta)^{n-t}d\theta}$; this is the Beta$(t+1, n-t+1)$ distribution. We find the peak of this distribution is the MLE, which is unsurprising since we started with a uniform distribution, so this new distribution is simply the likelihood function [I think]

b) $X_1, \ldots, X_n \sim N(\mu, 1)$. Prior distribution for μ given by $p(\mu) \sim N(0, \tau^{-2})$. $P(\mu \mid x_1 \ldots x_n) \propto f(x_1 \ldots x_n)p(\mu) \propto e^{-\frac{1}{2}\sum(x_i-\mu)^2}e^{-\frac{1}{2}\tau^2\mu^2}$; we could integrate at this stage but it would be very messy; instead we rearrange this as being $\propto e^{-\frac{1}{2}(n+\tau^2)(\mu-\frac{\sum x_i}{n+\tau^2})^2}$ which we can then recognise as $p(\mu \mid x_1 \ldots x_n) \sim N(\frac{\sum x_i}{n+\tau^2}, \frac{1}{\mu+\tau^2})$

c) $X_1, \ldots, X_n \sim \exp(\lambda)$ i.i.d; prior $\lambda \sim \exp(\mu)$. $P(\lambda \mid x_1 \ldots x_n) \propto (\prod_1^n \lambda e^{-\lambda x_i})\mu e^{-\lambda\mu} = \propto \lambda^n e^{-\lambda(\mu+\sum x_i)}$; we recognise this as being $\propto$ gamma$(n+1, \mu+\sum x_i)$ since $\Gamma(n,\theta)$ has $\frac{\theta^n t^{n-1}e^{-t}}{(n-1)!}$, so $p(\lambda \mid x_1 \ldots x_n) = \frac{\lambda^n(\mu+\sum x_i)^{n+1}e^{-\lambda(\mu+\sum x_i)}}{n!}$

All these give us distributions for $P(\theta \mid \text{data})$. Rather than simply taking the peak value, we might like to choose $\hat{\theta}$ to minimize some loss function.

5.3 Estimation within Bayesian Statistics

a) Say we have some loss function $L(\theta, a)$ between the true value θ and our estimate a, e.g. $(a-\theta)^2$. We want ti minimise $EL(\theta)$ (the expectation being taken wrt the posterior distribution) over a; it is $\int L(\theta,a)p(\theta \mid x_1 \ldots x_n)d\theta = \int(\theta-a)^2 p(\theta \mid x_1 \ldots x_n)d\theta$; to minimize we differentiate wrt a and put $0 = 2\int(a-\theta)p(\theta \mid x_1 \ldots x_n)d\theta$, so $a = \int \theta p(\theta \mid x_1 \ldots x_n)d\theta$, the posterior mean, so we take $\hat{\theta}$ to be this

b) $L(\theta, a) = |a-\theta|$ has $EL(\theta,a) = \int_{-\infty}^a (a-\theta)p(\theta \mid x_1 \ldots x_n)d\theta + \int_a^\infty p(\theta \mid x_1 \ldots x_n)d\theta$; differentiating and putting $= 0$ we have $0 = \int_{-\infty}^a p(\theta \mid x)d\theta - \int_a^\infty p(\theta \mid x)d\theta$, so a should be the median of the posterior distribution of θ

Example 5.4

$X_1, \ldots, X_n \sim P(\lambda), \lambda \sim \exp(1)$ i.e. $p(\lambda) = e^{-\lambda}, \lambda \geq 0$. $P(\lambda \mid x_1 \ldots x_n) \propto (\prod_1^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!})e^{-\lambda} \propto e^{-\lambda(n+1)}\lambda^{\sum x_I}$, the distribution of $\Gamma(\sum x_i + 1, n+1)$. The mean of this is $\frac{\sum x_i+1}{n+1}$, so this is our Bayes estimate for λ under quadratic loss (with this particular prior distribution). There is no neat expression for our estimate under absolute error loss; $\tilde{\lambda}$ is simply the value such that $\int_0^{\tilde{\lambda}} \frac{e^{-\lambda(n+1)}\lambda^{\sum x_i}(n+1)^{\sum x_i+1}}{(\sum x_i)!}d\lambda = \frac{1}{2}$.

6 Hypothesis Testing

6.1 The Neyman-Dearson framework

Say we have $X_1 \ldots X_n \sim f(\mid \theta)$ i.i.d. For estimation we use some $\hat{\theta}(x)$, for hypothesis testing we test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. H_0 is called the null hypothesis, H_1 is the alternate hypothesis. We could also have hpyotheses like $H_0 : f = f_0, H_1 : f = f_1$ or $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset, \Theta_0 \cup \Theta_0 = \Theta$ the entire paramater space. For now we consider $H_0 : f = f_0, H_1 : f \neq f_0$, a goodness-of-fit test.

6.2 Terminology

A simple hypothesis specifies f completely, e.g. $\theta = \theta_0$, wheras $\theta > \theta_0$ or $\theta \in \Theta_0$ would be a composite hypothesis. We will have some critical region C; we reject H_0 iff our data $x = (x_1 \ldots x_n) \in C \subset \mathbb{R}^n$. There are two types of errors: a type I error is rejecting H_0 when it is true, a type II error is not rejecting (which we may wish to distinguish from accepting) H_0 when it is false. Generally type I errors are "worse", e.g. H_0 = defendant is innocent in a murder case.

$P(\text{type I error}) = \alpha$ should be small; we generally set $\alpha = 0.01$ or 0.05. Let $P(\text{type II error}) = \beta$; we then have an optimization problem, to minimize β subject to fixed α. This α is called the size or significance level.

$\alpha = P(x \in C \mid H_0)$; for a simple $H_0 : \theta = \theta_0, \alpha = P(x \in C \mid \theta = \theta_0$, while for a composite $H_0 : \theta \in \Theta_0$, the size is $\sup_{\theta_0 \in \Theta_0} P(x \in C \mid \theta = \theta_0)$.

We considered likelihood $f(x_1 \dots x_n \mid \theta)$ as a function of θ; when testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ we consider the likelihood ratio $\frac{f(x_1 \dots x_n|\theta=\theta_1)}{f(x_1 \dots x_n|\theta=\theta_0)}$ and reject H_0 when this is large. For composite hypotheses we use $L_x(H_0, H_1) = \frac{L_x(H_1)}{L_x(H_0)}$ where $L_x(H_i) = \sup_{\theta \in \Theta_i} f(x_1 \dots x_n \mid \theta)$.

6.3 Likelihood ratio tests

$C = \{x : L_x(H_0, H_1) \geq k\}$ say. This gives us a likelihood ratio test.

Lemma 6.3 (Neyman-Pearson Lemma)

Say we have $H_0 : f = f_0$ to be tested against $H_1 : f = f_1$. Assume $f_1, f_0 > 0$ on the same regions and are continuous. Then amongst all test of size $\leq \alpha$ the test with the smallest probability of a type II error is given by $C = \{x : \frac{f_1(x)}{f_0(x)} \geq k\}$ where k is chosen such that $\alpha = P(X \in C \mid H_0) = \int_{x \in C} f_0(x)dx$ $[= \int_{\mathbb{R}^n} \phi_C(x) f_0(x)dx$; see below]. This is a popular tripos question.

Consider any test with size $\leq \alpha$; let its critical region be D. Let $\phi_D(x)$ be the indicator that $x \in D$. Then $0 \leq (\phi_C(x) - \phi_D(x))(f_1(x) - kf_0(x))$ by the definition of $\phi_C(x)$. Integrating over $\mathbb{R}^n$, $0 \leq P(X \in C \mid H_1) - kP(X \in C \mid H_0) - P(X \in D \mid H_1) + kP(X \in D \mid H_0) \leq P(X \in C \mid H_1) - P(X \in D \mid H_1 + \alpha - \alpha$ so $1 - P(X \in C \mid H_1) \leq 1 - P(X \in D \mid H_1)$ i.e. $P_C(\text{type II error}) \leq P_D(\text{type II error})$.

6.4 Single sample test mean with simple alternate, normal distribution with known variance

$x_1 \dots x_n \sim N(\mu, \sigma^2), \sigma^2$ known. $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$. $\frac{f(x|\mu_1 \sigma^2}{f(x|\mu_0, \sigma^2} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum(x_i - \mu_1)^2}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2}} = e^{\frac{\sum(x_i - \mu_0)^2 - \sum(x_i - \mu_1)^2}{2\sigma^2}}$. Assume $\mu_1 > \mu_0$, then this is monotone increasing in $\bar{x}$, so $\geq k$ iff $\bar{x} \geq$ some c. There is no need to compute the relationship between k and c and doing so would waste a lot of time; from now on we work purely with c; $C = \{x : \bar{x} \geq c\}$ some c. Recall $\bar{x} \sim N(\mu_0, \frac{\sigma^2}{n})$ if H_0 true. Let $Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \sim N(0, 1)$. If $\alpha = 0.05$ then $P_{H_0}(\bar{x} \geq c) \equiv P_{H_0}(Z \geq c') = 0.05 \Rightarrow c' = 1.645 \Rightarrow c = \mu_0 + \frac{\sigma 1.645}{\sqrt{n}}$.

Say we were testing $H_0 : \mu = 5$ against $H_1 : \mu = 6$ with $\sigma^2 = 1$ and have data $x = (5.1, 5.5, 4.9, 5.3)$. $\bar{X} = 5.2 \therefore Z = \frac{2(5.2-5)}{1} = 0.4 < 1.645$ so we don't reject H_0. However, notice that if we were testing $H_0 : \mu = 6$ against $H_1 : \mu = 5$ we have $Z = \frac{2(5.2-6)}{1} = -1.6 > -1.645$, so we don't reject H_0 in this case either.

7 Further aspects of Hypothesis Testing

α =size=significance level=P(reject $H_0 \mid H_0$) = P(type I error). $Z = 0.4$ in the example above. Rather than fixing $\alpha = \sup_{\theta \in \Theta_0} P(X \in C \mid \theta)$ and then testing H_0 against this, we can define the p-value $p^\star = \sup_{\theta \in \Theta_0} P_\theta(L_X(H_0, H_1) \geq L_x(H_0, H_1))$, the probability of obtaining a more extreme result - this is the smallest α such that we would reject H_0 if we were conducting a test of size α.

7.1 The power of a test

A type II error is not rejecting H_0 when it is false. For $\theta \in \Theta$ we define the power function $w(\theta) = P(X \in C \mid \theta)$; for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ this is $1 - \beta$. This is generally an increasing curve passing through α at $\theta = \theta_0$; of course it is easier to make the test if θ_1 is larger.

$w(\theta) = 1 - P$(type II error $\mid \theta$) for $\theta \neq \theta_0$.

7.2 Uniformly most powerful test

This is a difficult section: for $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ we know the best test is to reject H_0 if $Z \geq c$ some c (i.e. $\bar{X} \geq c'$). We notice that c is independent of the value of μ_1, so in fact we have the same test for any $\mu_1 > \mu_0$. We say this test is uniformly most powerful for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$; α is now $\sup_{\mu \leq \mu_0} P(x \in C \mid \mu)$ which we find $= P(X \in C \mid \mu = \mu_0)$ as before.

Example

$X_1 \dots X_n \sim N(\mu, \sigma^2)$, μ known. Test $H_0 : \sigma^2 \leq 1$ against $H_1 : \sigma^2 > 1$; to do this we first consider $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 = \sigma_1^2$, $\sigma_0^2 \leq 1 < \sigma_1^2$. $\frac{f(x|\mu\sigma_1^2)}{f(x|\mu\sigma_0^2)} = \frac{\sqrt{2\pi\sigma_1^2} e^{-\frac{1}{2\sigma_1^2}\sum(X_i-\mu)^2}}{\sqrt{2\pi\sigma_0^2} e^{-\frac{1}{2\sigma_0^2}\sum(X_i-\mu)^2}} = \left(\frac{\sigma_1}{\sigma_0}\right)^n e^{(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2})\sum(X_i-\mu)^2}$, an increasing function of $\sum(X_i - \mu)^2$. So we should reject H_0 if $\sum(X_i - \mu) \geq c$ for some c.

$X_i - \mu \sim N(0, \sigma^2)$ so $\sup_{\sigma_0^2 \leq 1} P(\sum(X_i - \mu)^2 \geq c \mid \sigma^2 = \sigma_0^2) = P(\sum(X_i - \mu)^2 \geq c \mid \sigma_0^2 = 1)$ and (if H_0 true) $X_i - \mu \sim N(0, 1)$. The sum of n such X_i is defined to be distributed as χ_n^2, the chi-squared distribution. We reject H_0 if $\sum(X_i - \mu)^2 \geq F_\alpha^{(n)}$ where $F_\alpha^{(n)}$ is the value at which the area above it under a χ_n^2 distribution is α; this does not depend on σ_1^2 so is our best test for $H_0 : \sigma^2 \leq 1$ against $H_1 : \sigma^2 > 1$.

7.3 Confidence intervals and hypothesis tests

Theorem 7.3

Suppose that for every θ_0 there is a test of size α of $H_0 : \theta = \theta_0$ against some H_1. Denote the acceptance region (i.e. complement of the critical region) of this by $A(\theta_0)$. Let $I(X) = \{\theta : X \in A \mid \theta\}$. This is a $100(1-\alpha)\%$ CI for θ, and conversely (i.e. if we have a CI we can form such a test for any θ). [lol θ, θ_0], as $P(X \in A(\theta_0) \mid \theta = \theta_0) = P(\theta \in I(X) \mid \theta = \theta_0) = 1 - \alpha$, $X \in A(\theta_0) \Leftrightarrow \theta \in I(X)$. So finding a 95% CI for μ and testing whether or not μ_0 is in this interval is equivalent to testing H_0 against H_1 at $\alpha = 0.05$.

Until now we have mostly covered one-tailed tests. We have a two-tailed test if we test a hypothesis such as $H_1 : \theta \neq \theta_0$; there are two possibile ways this can be true, namely $\theta < \theta_0$ and $\theta > \theta_0$. We generally arrange things such that if H_0 is true there is a probability $\frac{\alpha}{2}$ of a result in each tail.

7.4 The Bayesian perspective on hypothesis testing

Say we are tosing a coin and testing $H_0 : p = \frac{1}{2}$ against $H_1 : p > \frac{1}{2}$. One possible experiment is to toss it 5 times and count the number of heads; say we get HHHHT, then our p-value is $P(\text{number of heads} \geq 4 \mid H_0) = (\frac{1}{2})^5 + 5\frac{1}{2}(\frac{1}{2})^4 = 0.1875$. An alternative experiment is to toss the coin until we get our first tail; say we get HHHHT, then our p-value is the probability our first tail is on the fifth or later test, i.e. the probability of four heads, so $(\frac{1}{2}^4) = 0.625$. Thus if our α were 0.10, we would accept H_0 for the first experiment but reject it for the second. This is rather odd, since "the coin didn't know" which experiment we were doing. The Bayesian approach would give us the same answer in both cases, since $\frac{P(H_1|x)}{P(H_0|x)} = \frac{P(x|H_1)P(H_1)}{P(x|H_0)P(H_0)} = L_x(H_0H_1)\frac{P(H_1)}{P(H_0)}$ and $L_x(H_0H_1), \frac{p^4(1-p)}{(\frac{1}{2})^5}$ in this case, is independent of the choice of experiment.

8 Generalized likelihood ratio tests

8.1 χ^2 distribution

$X_1^2 + \cdots + X_n^2$ has χ_n^2 distribution when the X_i are i.i.d. $N(0,1)$ random variables. $\chi_n^2 \equiv \Gamma(\frac{1}{2}n, \frac{1}{2})$ [or possibly $\frac{1}{2n}$, lol lecturer]. The p.d.f. is $f(t) = \frac{(\frac{1}{2})^{\frac{n}{2}} t^{\frac{n}{2}-1} e^{-\frac{t}{2}}}{\Gamma(\frac{n}{2})}$, for $t > 0$. $\int_0^\infty (\frac{1}{2})^{\frac{n}{2}} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt = \Gamma(\frac{n}{2})$, so this is correctly normalized. See the fact (from IA probability) that for $X_1, X_2 \sim N(0,1)$, if we let $r^2 = X_1^2 + X_2^2$ then $r \sim$ exp.

8.2 Generalized likelihood ratio tests

$C = \{x : L_x(H_0, H_1) > k\}$. We generally test whether some $T(X)$ is $>$ some c; we need to find $P_{H_0}(T(X) > c)$. This is easy for e.g. the normal case with $T = \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, but how do we find it for more complicated distributions?

Say we are testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ with the $\Theta_i \subset \Theta = \{\theta = (\theta_1, \ldots, \theta_k)\}$. If we have H_0 of the form $\theta_{i_1} = \alpha_1, \ldots, \theta_{i_p} = \alpha_p$ for fixed $\alpha_1, \ldots, \alpha_p$, or $A\theta = b$ for some fixed $p \times k$ matrix A and p-vector b, or $\theta_i = \theta_i(\phi_i, \ldots, \phi_{k-p}) \forall i$. In all of these cases there are $k - p$ degrees of freedom.

Theorem 8.1

Suppose $\Theta_0 \subset \Theta_1$ and $|\Theta_1| - |\Theta_0| = p$ (where $|\Theta_i|$ denotes the number of degrees of freedom). Under certain conditions (which will not be stated for here, but hold for all usual cases) for $X_1, \ldots, X_n$ i.i.d. [and possibly only in the limit as $n \to \infty$ - lol lecturer] $2 \log L_X(H_0H_1) \sim \chi_p^2$ if H_0 is true, and $2 \log L_X$ is larger if H_0 is not true. So we reject H_0 if $2 \log L_X(H_0H_1) > c$ where $\alpha = P(\omega > c)$ for $\omega \sim \chi_p^2$. We shall not proove this here.

Lemma 8.2

For $X_1, \ldots, X_n$ i.i.e. as $N(\mu, \sigma^2)$:

$\max_\mu f(x \mid \mu\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{\sum(X_i-\bar{X})^2}{2\sigma^2}}$

$\max_{\sigma^2} f(x \mid \mu\sigma^2) = (2\pi\frac{\sum(X_i-\mu)^2}{n})^{-\frac{n}{2}} e^{-\frac{n}{2}}$

$\max_{\mu,\sigma^2} f(x \mid \mu\sigma^2) = (2\pi\frac{\sum(X_i-\bar{X})^2}{n})^{-\frac{n}{2}} e^{-\frac{n}{2}}$

8.3 Single sample, known variance

Test $H_0 : \mu = \mu_0$ against $H_1 : \mu$ arbitrary (in practice, $\mu \neq \mu_0$) [I think, lol lecturer]. $L_X(H_0H_1) = \frac{\sup_\mu f(x|\mu\sigma^2)}{f(x|\mu_0\sigma^2)} = e^{\frac{1}{2\sigma^2}n(\bar{X}-\mu)^2} \therefore 2\log L_x = \frac{1}{\sigma^2}n(\bar{X}-\mu)^2 = z^2$ where $z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \chi_1^2$ if H_0 is true. The one-tailed χ^2 test we perform is equivalent to a two-tailed test for a normal distribution.

8.4 Single sample, test variance, known mean

Say $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$. $L_X(H_0H_1) = \frac{\sup_{\sigma^2} f(x|\mu\sigma^2)}{f(x|\mu\sigma_0^2)}$; we find $2\log L_X$ [L_X is short for $L_X(H_0H_1)$] is $n(t - 1 - \log t)$ where $t = \frac{\sum(X_1-\mu)^2}{n\sigma_0}$. If H_0 is true then $\frac{\sum(X_i-\mu)^2}{n\sigma_0^2} \sim \chi_n^2$; this is unsurprising since each $\frac{X_i-\mu}{\sigma} \sim N(0,1)$ [I think; lecturer was on really bad form this lecture]

8.5 Two samples, test equality of means, known common variance

Say $X_1, \ldots, X_m$ i.i.d as $N(\mu_1, \sigma^2), Y_1, \ldots, Y_n$ i.i.d. as $N(\mu_2, \sigma^2)$. We test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$; $L_X(H_0H_1) = \frac{\sup_{\mu_1,\mu_2} f(X|\mu_1\sigma^2)f(Y|\mu_2\sigma^2)}{\sup_\mu f(X|\mu\sigma^2)f(Y|\sum\sigma^2)}$, which we can find to be $e^{\frac{1}{2\sigma^2}\frac{mn}{m+n}(\bar{X}-\bar{Y})^2}$; $2\log L_X \sim \chi_1^2$ since $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{m}), \bar{Y} \sim N(\mu_2, \frac{\sigma^2}{n})$ so if H_0 true $\bar{X} - \bar{Y} \sim N(0, \sigma^2(\frac{1}{m} + \frac{1}{n}))$ so $Z = (\bar{X} - \bar{Y})\frac{1}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim N(0,1)$ and so $Z^2 = \frac{(\bar{X}-\bar{Y})^2}{\sigma^2}\frac{mn}{m+n} \sim \chi_1^2$.

8.6 Goodness of fit test

Say we have k categories of possible results with respective probabilities p_i, and obtain a result of x_i in each category with $\sum x_i = n$ (of course $\sum p_i = 1$). We test $H_0 : p_i = p_i(\theta)$ some $\theta \in \Theta_0$ against $H_1 : p_i$ unrestricted, e.g. $H_0 : p_i = \binom{k}{i}\theta^i(1-\theta)^{k-i}$. $P(X_1 \ldots X_k \mid p_1 \ldots p_k) = \frac{n!}{x_1!\ldots x_k!}p_1^{x_1} \ldots p_k^{x_k}$, so $\sup_{H_1} \log f(x) =$ some constant $+ \sup\{\sum x_i \log p_i \mid 0 \leq p_i \leq 1, \sum p_i = 1\}$. Using Lagrangian multipliers as per the Optimisation course we find $\hat{p}_i = \frac{X_i}{n}$. $\sup_{H_0} \log f(x) =$ constant $+ \sup_\theta\{\sum x_i \log p_i(\theta)\}$; we reject H_0 if $2\log L_x(H_0, H_1)$ is large. $H_0 : p_i = p_i(\theta), \theta \in \Theta_0$ has $|\Theta_0| = p$ degrees of freedom, while $H_1 : p_i$ arbitrary has $k - 1$ degrees of freedom (since we still have the constraint that $\sum_1^k p_i = 1$), so we test against $\chi^2_{k=1-p}$, the number of degrees of freedom being the number of boxes - the number of paramaters estimated (for H_0) - 1.

[Note: I have sometimes used $l_X(H_0H_1)$ above for the likelihood ratio; in lectures $L_X(H_0H_1)$ was always used]

9 Chi-squared tests of categorical data

9.1 Pearson's chi-squared test

We saw for $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ with $\Theta_0 \subset \Theta_1$, $2 \log L_X(H_0 H_1) \sim \chi^2_{|\Theta_1|-|\Theta_0|}$ [if H_0 true]; if there are k possible outcomes with probability p_i of each and we obtain x_i outcomes of type i from $\sum x_i = n$ trials, and test $H_1 : p_1, \dots, p_k$ anything, $\sum p_i = 1, 0 \le p_i \le 1$ against $H_0 : p_i = p_i(\theta)$ e.g. $p_i = \binom{n}{i}\theta^i(1-\theta)^{n-i}$ or $\frac{\theta^i e^{-\theta}}{i!}$: For $\sup_{H_1} P(X_1 \dots X_K \mid p_1 \dots p_k) = \sup_{H_1} \frac{n!}{X_1! \dots X_n!}$ we use Lagrangian multipliers and maximise $L = \log P(\dots) + \lambda(1 - \sum p_i)$; $\frac{\partial L}{\partial p_i} = \frac{x_i}{p_i} - \lambda = 0 \therefore \hat{p}_i \propto x_i \therefore \hat{p}_i = \frac{x_i}{n}$. $\sup_{H_0} p(X_1 \dots X_n \mid p_1(\theta) \dots p_n(\theta))$ is really a sup over θ. Then $2 \log L_X(H_0 H_1) = 2 \sup_{p_1 \dots p_k} \log P(X \mid H_1) - 2 \sup_\theta \log P(X \mid H_0) = 2 \sum x_i \log(\frac{\frac{x_i}{n}}{p_i(\hat{\theta})}$. Pearson wanted to simplify this; let $o_i = x_i$ the observed number in the ith cell, $e_i = np_1(\hat{\theta})$, the expected number in the ith cell if H_0 is true, and $\delta_i = o_i - e_i$. Then the above becomes $2 \sum o_i \log \frac{o_i}{e_i} = 2 \sum_i (e_i + \delta_i) \log(1 + \frac{\delta_i}{e_i}) = 2 \sum_i (e_i + \delta_i)(\frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2} + \dots)$ (δ_i will be small if H_0 is true) which is $2(\sum \delta_i + \sum \frac{\delta_i^2}{e_i} - \sum \frac{\delta_i^2}{2e_i} + \dots) = 2 \sum \frac{\delta_i^2}{2e_i}$ since $\sum \delta_i = \sum o_i - \sum e_i = n - n = 0$, and this is $\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$, Pearson's chi-squared statistic. This is $\sim \chi^2_{k-1-p}$ where $p = |\Theta_0|$ or equivalently the number of paramaters we estimate to fit the null hypothesis to the data.

Observe that $\sum \frac{(o_i - e_i)^2}{e_i} = \sum \frac{o_i^2}{e_i} - 2 \sum o_i + \sum e_i = \sum \frac{o_i^2}{e_i} - n$ (since $\sum e_i = n = \sum o_i$).

9.2 χ^2 test of homogeneity

Say we have a table of results, e.g. columns of whether a patient survived against rows of whether they were male or female, and we want to test $H_0 : p_{ij} = p_j$, i.e. the distribution of each row is the same, agaists H_1 that the p_{ij} (the probability of getting result j for a result in the ith row) are arbitrary (such that the $\sum_j p_{ij} = 1$ for each i); the details are in the printed notes for this course.

For H_0, $\hat{p}_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}$ where $x_{\cdot j} = \sum_i x_{ij}$ and similarly, and for H_1, $\hat{p}_{ij} = \frac{x_{ij}}{x_{i\cdot}}$, so $2 \log L_x(H_0 H_1) = 2 \sum_i \sum_j x_{ij} \log(\frac{x_{ij} x_{\cdot\cdot}}{x_{\cdot j} x_{i\cdot}})$; $o_{ij} = x_{ij}, e_{ij} = x_{i\cdot} \hat{p}_j = \frac{x_{i\cdot} x_{\cdot j}}{x_{\cdot\cdot}}$ so this is $2 \sum_i \sum_j o_{ij} \log(\frac{o_{ij}}{e_{ij}})$, which as before is approximately $\sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$. This approximation is in general valid for $e_{ij} \ge 5 \forall i, j$; this is useful for e.g. knowing where to "truncate" the cells for a poisson distribution (since we must have a cell $\ge n$ for some n if we want to have a finite number of cells). $H_0 : p_{ij} = p_j$ has $n - 1$ degrees of freedom where n is the number of columns, and H_1 has $m(n-1)$ degrees of freedom where m is the number of rows, so the number of degrees of freedom to use in the test is $m(n-1) - (n-1) = (m-1)(n-1)$.

9.3 χ^2 test of row column independence, contingency tables

Say we have a similar table of results, but this time want to test $H_0 : p_{ij} = p_i q_j$ i.e. rows and columns are independent, against H_1 that the p_{ij} are arbitrary. $|\Theta_1| = mn - 1$ [O RLY] and $|\Theta_0| = m - 1 + n - 1$ so the number of degrees

of freedom for the test is $mn-1-m+1-n+1=(m-1)(n-1)$; $\hat{p}_i = \frac{x_{i\cdot}}{x_{\cdot\cdot}}, \hat{q}_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}, \hat{p}_{ij} = \frac{x_{ij}}{x_{\cdot\cdot}}, e_ij = \hat{p}_i\hat{q}_j x_{\cdot\cdot}$; we find ourselves considering $\log(\frac{\hat{p}_{ij}}{\hat{p}_i\hat{q}_j}) = \log(\frac{x_{ij}x_{\cdot\cdot}}{x_{\cdot j}x_{i\cdot}}) = \log\frac{o_{ij}}{e_{ij}}$ and we have the exact same analysis as in 9.2; we test $T=\sum\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ against a $\chi^2_{(m-1)(n-1)}$ distribution, and the e_{ij} are exactly the same for both analyses; the above $x_{i\cdot}\hat{p}_j = \frac{x_{i\cdot}x_{\cdot j}}{x_{\cdot\cdot}} = x_{\cdot\cdot}\hat{p}_i\hat{p}_j$ in this test. This is interesting, since the origins of the two tests are philisophically quite different.

10 Distributions of the sample mean and variance

10.1 Simpson's paradox

Lulz. Like the one you won 10 for explaining.

10.2 Transformations of variables

For $X_1 \dots X_n$ i.i.d. as $N(\mu,\sigma^2)$, let $\sum(X_i-\bar{X})^2 = S_{XX}$. Say $X_i = x_i(Y_1,\dots,Y_n)\forall i$, then $f_Y(y_1,\dots,y_n) = f_X(x_1(y),\dots,x_n(y))\left|\begin{array}{ccc}\frac{\partial x_1}{\partial y_1} & \dots & \frac{\partial x_1}{\partial y_n}\\ \dots & & \dots \\ \frac{\partial x_n}{\partial y_1} & \dots & \frac{\partial x_n}{\partial y_n}\end{array}\right|$, where the matrix is the Jacobian; compare this with a change of variables in integration when e.g. $dxdy = rdrd\theta$.

Example 10.2

Suppose $X_1 \sim \Gamma(n_1,\lambda), X_2 \sim \Gamma(n_2\lambda)$ independent and let $Y_1 = \frac{X_1}{X_1+X_2}, Y_2 = X_1+X_2$. $f_X(x_1x_2) = \frac{\lambda^{n_1}e^{-\lambda x_1}x_1^{n_1-1}}{(n_1-1)!}\frac{\lambda^{n_2}e^{-\lambda x_2}x_2^{n_2-1}}{(n_2-1)!}$; $x_1 = y_1y_2, x_2 = y_2-y_1y_2$ so $J(y_1y_2) = \left|\begin{array}{cc} y_2 & y_1 \\ -y_2 & 1-y_1\end{array}\right| = y_2$ so $f_Y(y_1y_2 = \frac{\lambda^{n_1+n_2}(y_1y_2)^{n_1-1}(y_2-y_1y_2)^{n_2-1}e^{-\lambda y_2}}{(n_1-1)!(n_2-1)!}\times y_2 = \frac{(n_1+n_2-1)!}{(n_1-1)!(n_2-1)!}y_1^{n_1-1}(1-y_1)^{n_2-1} \times \frac{\gamma^{n_1+n_2}e^{-\lambda\gamma_2}y_2^{n_1+n_2-1}}{(n_1+n_2-1)!}$ so Y_1, Y_2 are independent and distributed as $\beta(n_1,n_2), \Gamma(n_1+n_2,\lambda)$ respectively.

10.3 Orthogonal transformations of normal random variables [section 10.2 in lectures]

Lemma 10.3

Let $X_1,\dots,X_n$ be independent random variables, distributed as $N(\mu_i,\sigma^2)$ respectively, $A=(a_{ij})$ an orthogonal matrix and $Y=AX$; this is a vector of independently distributed components with each $Y_i \sim N((A\mu)_i,\sigma^2)$: $f_X(x_1,\dots,x_n \mid \mu\sigma^2) = \prod f(x_i \mid \mu_i\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}(X-\mu)^T(X-\mu)} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2}$. Now $Y=AX \therefore X = A^TY \therefore \frac{\partial X_i}{\partial Y_j} = a_{ji} \therefore J(y_1,\dots,y_n) = |\det A^T| = 1$, so $f_Y(y_1\dots y_n \mid \mu\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{(A^Ty-\mu)^T(A^Ty-\mu)}{2\sigma^2}}\times 1 = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{(A^Ty-A^TA\mu)^T(A^Ty-A^TA\mu)}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{(y-A\mu)^T(y-A\mu)}{2\sigma^2}} \Rightarrow Y_1,\dots,Y_n \sim N(A\mu,\sigma^2I)$ a multivariate normal dis-

tribution, or equivalently the $Y_i \sim N((A\mu)_i, \sigma^2)$ independently; we can also proove this result via moment generating functions.

10.4 The distributions of $\bar{X}, S_{XX}$

Lemma 10.4

Let $X_1, \ldots, X_n$ be i.i.d. as $N(\mu\sigma^2)$ and $\bar{X} = \frac{1}{n}\sum_1^n X_i, S_{XX} = \sum(X_i - \bar{X})^2$. Then:

i) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n}), n(\bar{X} - \mu)^2 \sim \sigma^2\chi_1^2$

ii) $X_i - \mu \sim N(0, \sigma^2), \sum(X_i - \mu)^2 \sim \sigma^2\chi_n^2$

iii) $\sum(X_i - \mu)^2 = S_{XX} + n\bar{X} - \mu)^2$

iv) $\frac{S_{XX}}{n-1}$ is an unbiased estimator for σ^2

v) $\bar{X}, S_{XX}$ are independent random variables

vi) $S_{XX} \sim \sigma^2\chi_{n-1}^2$

i) and ii) are immediate, iii) comes from consideration of $\sum(X_i - \bar{X} + \bar{X} - \mu)^2$ (the cross product terms of which are 0 since $\sum X_i = n\bar{X}$); iv) we have already prooven.

For v), let $Y = A(X - \mu) = (Y_1 = \sqrt{n}(\bar{X} - \mu), Y_2, \ldots, Y_n)$; we can choose $A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ & \cdots & \cdots & \end{pmatrix}$ to be orthogonal. We know $Y_1, \ldots, Y_n$ are independent normal random variables, so $Y_1 = \sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2)$ and is independent of $Y_2, \ldots, Y_n$. $\sum_2^n Y_i^2 = \sum_1^n (X_i - \mu)^2 - Y_1^2$ since $Y^TY = (X - \mu)^T A^T A(X - \mu) = (X - \mu)^T(X - \mu)$ so $\sum_1^n Y_i^2 = \sum_1^n (X_i - \mu)^2$ and so since $y_1^2 = n(\bar{X} - \mu)^2$ this is S_{XX} so S_{XX}, Y_1 are independent, i.e. $S_{XX}, \bar{X}$ are independent. $Y_i \sim N(0, \sigma^2)$ for $i = 2, \ldots, n$ and these are independent so $Y_2^2 + \cdots + Y_n^2 \sim \sigma^2\chi_{n-1}^2$ (and we have vi)).

10.5 Student's t-distribution

To test $H_0 : \mu = \mu_0$ for $X_1, \ldots, X_n$ i.i.d. as $N(\mu, \sigma^2)$ for σ^2 known we would reject H_0 if $\left|\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}\right|$ is large. To perform this test for unknown σ^2 we use $\hat{\sigma}^2 = \frac{S_{XX}}{n-1}$; we discover $\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{S_{XX}}{n-1}}}$ has the t_{n-1} distribution independent of the true value of σ^2; informally $t_{n-1} \equiv \frac{N(0,1)}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}}$. We could calculate the distribution function but it is messy; it "looks like a spread out normal", and $t_n \to N(0, 1)$ as $n \to \infty$. For $X \sim N(0, 1)$ and $W \sim t_{n-1}$, $P(X > t) < P(w > t) \forall t$; the RHS is decreasing in increasing n [???].

11 The t-test

Take $X_1, \ldots, X_n$ i.i.d. as $N(\mu\sigma^2)$ unless otherwise stated.

11.1 Confidence interval for mean, unknown variance

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ or equivalently $\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma^2} \sim N(0,1)$; we saw above this is independent of $S_{XX} = \sum(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$ or equivalently $\frac{S_{XX}}{\sigma^2} \sim \chi^2_{n-1}$. σ is a "nuisance paramater"; we are not interested in its value but nevertheless have to consider it in our test.

We saw $\frac{(\bar{X}-\mu)\sqrt{n}}{\sqrt{\frac{S_{XX}}{n-1}}} \sim t_{n-1}$, so a CI of size $(1-a)100\%$ is is given by $1-\alpha = P(-t^{(n-1)}_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\frac{S_{XX}}{n-1}}} \leq t^{(n-1)}_{\frac{\alpha}{2}})$ [Yes, notation did just change at random with no explanation], i.e. $P(\bar{X} - t_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}})$ where $\hat{\sigma}^2 = \frac{S_{XX}}{n-1}$; compare this with the $\bar{X} \pm N_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ we would use when working with known σ.

11.2 Single sample, test mean, unknown variance

Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Then $L_X(H_0H_1) = \frac{\sup_{\mu\sigma^2} f(x|\mu\sigma^2)}{\sup_{\sigma^2} f(x|\mu_0\sigma^2} = \frac{\left(2\pi\sum\frac{(X_i-\mu_0)^2}{n}\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}}{\left(2\pi\sum\frac{(X_i-\bar{X})^2}{n}\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}} = \left(\frac{\sum(X_i-\bar{X})^2+n(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2}\right)^{\frac{n}{2}}$ since $X_i - \mu_0 = X_i - \bar{X} + \bar{X} - \mu_0$; this is $\left(1 + \frac{n(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2}\right)^{\frac{n}{2}}$; $\frac{n(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2} = T^2(n+1)$ where $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sqrt{\frac{S_{XX}}{n-1}}}$, and this is large when T is large. This makes sense, since if H_0 is true then $T \sim t_{n-1}$.

11.3 Two samples, test equality of means, unknown common variance

$X_1 \ldots X_m \sim N(\mu_1\sigma^2), Y_1 \ldots Y_n \sim N(\mu_2\sigma^2)$, test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. $L_X(H_0H_1) = \frac{\sup_{\mu_1\mu_2\sigma^2} f(x,|\mu_1\mu_2\sigma^2)}{\sup_{\mu\sigma^2} f(x,y|\mu\mu\sigma^2)}$; we find that we reject H_0 if $\frac{(\bar{X}-\bar{Y})^2}{S_{XX}+S_{YY}}$ is large. We can find this more "intuitively" by $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{m}), \bar{Y} \sim N(\mu_2, \frac{\sigma^2}{n})$ so $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{m} + \frac{1}{n}))$; if H_0 is true then $\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$. $S_{XX} \sim \sigma^2\chi^2_{m-1}, S_{YY} \sim \sigma^2\chi^2_{n-1}$ so $T = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{S_{XX}+S_{YY}}{m+n-2}}} \sim t_{m+n-2}$ So we reject if $|T| > t^{(n+m-2)}_{\frac{\alpha}{2}}$.

11.4 Single sample, test variance, mean unknown

We test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$; this time μ is our nuisance paramater. $L_X(H_0H_1) = \frac{\sup_{\mu\sigma^2} f(x|\mu\sigma^2)}{\sup_{\mu} f(x|\mu\sigma_0^2)}$ which we eventually find is large when $T = \frac{\sum(X_i-\bar{X})^2}{n\sigma_0^2}$ differs substantially from 1. We know $S_{XX} \sim \sigma_0^2\chi^2_{n-1}$ if H_0 is true; we want $P(\frac{S_{XX}}{\sigma_0^2} < a_1 \mid H_0) + P(\frac{S_{XX}}{\sigma_0^2} > a_2 \mid H_0) = \alpha$.

12 The F-test and analysis of variance

12.1 F-distribution

For $X \sim \chi^2_m$ we can see this as $X = \omega_1^2 + \cdots + \omega_m^2$ for ω_i i.i.d. as $N(0,1)$. $EX = m, \mathrm{Var}(X)$ we find to be $2m$. If $Y \sim \chi^2_n$ independently of X we say $Z = \frac{\frac{X}{m}}{\frac{Y}{n}} \sim F_{m,n}$; we clearly have $\frac{1}{Z} \sim F_{n,m}$.

12.2 Two samples, compare variance

Say $X_1, \ldots, X_n$ i.i.d. as $N(\mu_1, \sigma_1^2)$, $Y_1, \ldots, Y_n$ independently i.i.d. as $N(\mu_2, \sigma_2^2)$, and we want to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$. As always we use a LRT; $L_X(H_0 H_1) = \frac{\sup_{\sigma_1,\sigma_2,\mu_1,\mu_2} f(x,y|\mu_1,\sigma_1^2,\mu_2,\sigma_2^2}{\sup_{\sigma,\mu_1,\mu_2} f(x,y|\mu_1,\sigma^2,\mu_2,\sigma^2}$; we find we want to consider $\frac{S_{XX}}{S_{YY}}$. $\frac{S_{XX}}{m-1} = \hat{\sigma}_1$ and similarly for Y ; $S_{XX} = \sum_{i=1}^m (X_i - \bar{X})^2 \sim \sigma_1^2 \chi^2_{m-1}$, so $T = \frac{\frac{S_{XX}}{m-1}}{\frac{S_{YY}}{n-1}} \sim \frac{\sigma_1^2}{\sigma_2^2} F_{m-1,n-1}$; if H_0 is true this is $\sim F_{m-1,n-1}$ so we should reject H_0 if T lies in the lower or upper tail of such a distribution.

12.3 Non-central χ^2

For $X_i \sim N(\mu_i, \sigma^2)$, $X_1^2 + \cdots + X_n^2 \sim \chi^2_n(\lambda)$, a non-central χ^2_n distribution, where $\lambda = \mu_1^2 + \cdots + \mu_n^2$.

12.4 One-way analysis of variance (ANOVA)

This is used to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against H_1 that the μ_i are general, where we have n_i samples from each of k different populations; $X_{ij} = \mu_i + \epsilon_{ij}$ for $j = 1, \ldots, n_i, i = 1, \ldots, k$. Assume the ϵ_{ij} are IID as $N(0, \sigma^2)$ for σ^2 unknown. Then set $\bar{X}_{\cdot\cdot} = \frac{\sum_{ij} X_{ij}}{\sum_i n_i} = \hat{\mu}$ the "overall mean", $\bar{X}_{i\cdot} = \frac{\sum_j X_{ij}}{n_i} = \hat{\mu}_i$ the "sample mean". Then for $N = \sum_i n_i$, $L_X(H_0 H_1) = \frac{\sup_{\mu_1,\ldots,\mu_k,\sigma^2} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\sum_{ij} \frac{(X_{ij}-\mu_i)^2}{2\sigma^2}}}{\sup_{\mu,\sigma^2} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\sum_{ij} \frac{(X_{ij}-\mu)^2}{2\sigma^2}}}$ which we find is $\left(\frac{S_0}{S_1}\right)^{\frac{N}{2}}$ where $S_0 = \sum_{ij} (x_{ij} - \bar{x}_{\cdot\cdot})^2, S_1 = \sum_{ij} (X_{ij} - \bar{X}_{i\cdot})^2$; we have clearly $S_0 > S_1$, and $S_0 = \sum (x_{ij} - \bar{x}_{i\cdot} + \bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \sum (x_{ij} - \bar{x}_{i\cdot})^2 + \sum_i n_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$ (the cross terms are zero by summing over j before i); this is $S_1 + S_2$ where $S_2 = \sum_i n_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \sum_i n_i (\hat{\mu}_i - \hat{\mu})^2$, so $\frac{S_0}{S_1}$ is large when $\frac{S_2}{S_1}$ is large; if H_0 is true we can find $S_2 \sim \sigma^2 \chi^2_{k-1}, S_1 \sim \sigma^2 \chi^2_{N-k}$ independently; the distribution of S_1 coming from the fact that $\sum_j (X_{ij} - \bar{X}_{i\cdot})^2 \sim \sigma^2 \chi^2_{n_i - 1}$ independent of the μ_i. So we reject H_0 if $T = \frac{\frac{S_2}{k-1}}{\frac{S_1}{N-k}}$ is large compared to the $F_{k-1,N-k}$ distribution which it takes if H_0 is true.

13 Linear regression

[This lecture was missed]

14 Hypothesis tests in regression models

Say we have data $Y_1, \ldots, Y_n$ and assume $Y_i = \alpha + \beta w_i + \epsilon_i$ for unknown paramaters α, β with the ϵ_i i.i.d. as $N(0, \sigma^2)$ and the w_i known with $\sum w_i = 0$. We saw in the previous lecture than the MLEs are $\hat{\alpha} = \bar{Y}, \hat{\beta} = \frac{\sum Y_i w_i}{\sum w_i^2}$; we can write

$\vec{Y} = \begin{pmatrix} Y_1 \\ \ldots \\ Y_n \end{pmatrix} \sim N(\alpha\vec{1} + \beta\vec{w}, \sigma^2 I)$ a "multivariate normal" distribution.

14.1 Theorem

i) $\hat{\alpha} = \bar{Y} \sim N(\alpha, \frac{\sigma^2}{n})$

ii) $\hat{\beta} = \frac{S_{wY}}{S_{ww}} \sim N(\beta, \frac{\sigma^2}{\vec{w}^T\vec{w}})$ independent of $\hat{\alpha}$

iii) We say $S = \sum(Y_i - \alpha - \beta w_i)^2$ is minimised [by the above $\hat{\alpha}, \hat{\beta}$]by R, the residual sum of squares, which is $\sim \sigma^2\chi^2_{n-2}$ independent of $\hat{\alpha}, \hat{\beta}$. We can find $R = \sum Y_i^2 - n\hat{Y}^2 - (\vec{w}^T\vec{w})\hat{\beta}^2$.

iv) Therefore, $\hat{\sigma}^2 = \frac{R}{n-2}$

Recall the proof that $S_{XX}, \bar{X}$ are independent. Let $A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \ldots & \frac{1}{\sqrt{n}} \\ \sqrt{\vec{w}^T\vec{w}}w_1 & \sqrt{\vec{w}^T\vec{w}}w_2 & \ldots & \sqrt{\vec{w}^T\vec{w}}w_n \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$.

We can choose the remainder of the matrix such that $AA^T = I$ since we have that the inner product of the first row with itself is 1, likewise the second row, and the inner product of the first and second rows is 0 since $\sum w_i = 0$. So since $\vec{Y} \sim N(\alpha\vec{1} + \beta\vec{w}, \sigma^2 I)$, $\vec{Z} = A\vec{Y} \sim N(A(\alpha\vec{1} + \beta\vec{w}), \sigma^2 I)$ i.e. the components of Z are independent normal random variables with variance σ^2. We have $Z_1 = \sqrt{n}\hat{\alpha} = \sqrt{n}\bar{Y} \sim N(\sqrt{n}\alpha, \sigma^2)$ so i) holds, $Z_2 = \sqrt{\vec{w}^T\vec{w}}\hat{\beta} \sim N(\sqrt{\vec{w}^T\vec{w}}\beta, \sigma^2)$ so ii) holds. $Z_3, \ldots, Z_n$ are all $N(0, \sigma^2)$ independent - they have mean 0 since these are vectors orthogonal to $\vec{1}, \vec{w}$ so the relevant columns of $A(\alpha\vec{1} + \beta\vec{w}) = 0$. So $\sum_1^n Z_i^2 = n\bar{Y}^2 + (\vec{w}^T\vec{w})\hat{\beta}^2 + \sum_3^n Z_i^2 = \sum Y_i^2$ since $Z^T Z = Y^T Y$. $\sum Y_i^2 = \|Y - -\hat{\alpha}\vec{1} - \hat{\beta}\vec{w} + \hat{\alpha}\vec{1} + \hat{\beta}\vec{w}\|^2$ and the cross product terms can be found to be 0 so this $= \|Y - \hat{\alpha}\vec{1} - \hat{\beta}\vec{w}\|^2 + n\hat{\alpha}^2 + \hat{\beta}^2\|w\|^2 = R + n\bar{Y}^2 + (\vec{w}^T\vec{w})\hat{\beta}^2$ so $R + \sum_3^n Z_i^2 \sim \chi^2_{n-2}$ and we have iii) (and hence iv)).

14.2 Tests and CIs

To test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$: if H_0 is true $\hat{\beta} \sim N(\beta_0, \frac{\sigma^2}{\vec{w}^T\vec{w}}$ and so $\frac{(\hat{\beta} - \beta_0)\sqrt{w^T w}}{\sigma} \sim N(0, 1)$, and $\frac{R}{(n-2)\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}$ so $T = \frac{(\hat{\beta} - \beta_0)\sqrt{w^T w}}{\sqrt{\frac{R}{n-2}}} \sim t_{n-2}$ and we reject H_0 if $|T| > t^{(n-2)}_{\frac{\alpha}{2}}$.

To find a $(1-\alpha)100\%$ confidence interval for β we have $\frac{(\hat{\beta} - \beta)\sqrt{w^T w}}{\sqrt{\frac{R}{n-2}}} \sim t_{n-2}$ so $P(\hat{\beta} - t^{(n-2)}_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{w^T w}} \leq \beta \leq \hat{\beta} + t^{(n-2)}_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{w^T w}}) = 1 - \alpha$ where $\hat{\theta} = \sqrt{\frac{R}{n-2}}$.

A confidence interval for the value of Y that would be observed at a given w_0: $Y = \alpha + \beta w_0 + \epsilon_0 \sim N(\alpha + \beta w_0, \sigma^2)$. Let $\hat{Y} = \hat{\alpha} + \hat{\beta}w_0$. Then $Y - \hat{Y} \sim$

$N(0, \sigma^2(1 + \frac{1}{n} + \frac{w_0^2}{w^T w}))$ (since $\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n}, \mathrm{Var}(w_0\hat{\beta}) = \frac{w_0^2\sigma^2}{w^T w}$; note this is a "predictive confidence interval"; it is a CI for the value that would be measured at w_0 rather than the "true" value $\alpha+\beta w_0$ (to get one for that, we would remove the $\sigma^2 \times 1$ term from the variance)). So our $(1-\alpha)100\%$ confidence interval for Y is $\left[\hat{Y} - t_{\frac{\alpha}{2}}^{(n-2)}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{w_0^2}{w^T w}}, Y + t_{\frac{\alpha}{2}}^{(n-2)}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{w_0^2}{w^T w}}\right]$.

14.3 The correlation coefficient

The sample correlation coefficient of $(X_1, \ldots, X_n), (Y_1, \ldots, Y_n)$ is defined to be $R = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}$ (this is the commonly published correlation coefficient; 0 for no correlation, 1 for a perfect positive correlation, -1 for a perfect negative correlation). To test for a correlation between the two sets of data we let $w_i = X_i - \bar{X}$ and test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. We have $Y_i = \alpha + \epsilon_i$ and $T = \frac{\hat{\beta}\sqrt{S_{XX}}}{\sqrt{\frac{R}{n-2}}}$ which we find $= \frac{\frac{S_{XY}}{S_{XX}}\sqrt{S_{XX}\sqrt{n-2}}}{\sqrt{S_{YY}^2 - \frac{S_{XY}^2}{S_{XX}}}} = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$. So as we would expect we reject H_0 when r is far away from 0, i.e. close to ± 1 (since we always have $|r| \leq 1$.

$S_{YY} = \sum(Y_i - \bar{Y})^2$; $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ has $\sum(\hat{Y}_i - \hat{Y})^2 = \frac{S_{XY}^2}{S_{XX}}\frac{\sum(\hat{Y}_i-\hat{Y})^2}{\sum(Y_i-\hat{Y})^2} = r^2$.

14.4 Testing linearity

Say we have data $Y_{ij} = \alpha + \beta X_i + \epsilon_{ij}, j = 1, \ldots, m, i = 1, \ldots, n$, i.e. we have made m observations at each distinct point, for n different points. Let $\bar{Y}_i = \frac{1}{n}\sum_j Y_{ij}$, then $\bar{Y}_i = a + bX_i + \eta_i = \alpha + \beta(X_i - \bar{X}) + \eta_i$ with $\eta_i \sim N(0, \frac{\sigma^2}{m})$ independent of $\sum(Y_{ij} - \bar{Y}_i)^2 \sim \sigma^2\chi_{m-1}^2$ [$\sim$ not in notes but must be there for sanity]. We could do a linear regression analysis of $\bar{Y}_i$ on X_i; we get the residual sum of squares $\sum(\bar{Y}_i - \hat{\alpha} - \hat{\beta}(X_i - \bar{X}))^2 \sim \frac{\sigma^2}{m}\chi_{n-2}^2$. $F = \frac{\frac{m\sum(\bar{Y}_i-\hat{\alpha}-\hat{\beta}(X_i-\hat{\alpha})^2)}{n-2}}{\frac{\sum_{ij}(Y_{ij}-\bar{Y}_i)^2}{n(m-1)}}$, the ratio of variation explained by the linear regression to total variation, is $\sim F_{n-2,n(m-1)}$ and $\sum(Y_{ij} - \bar{Y})^2 \sim \sigma^2\chi_{nm-2}^2$.

14.5 Analysis of variance in regression models

Consider e.g. comparing rows k, l; $Y_{ij} = \alpha_i + \beta x_{ij}$ where i is which population the sample is taken from, j is the number of the sample; say the sample size is n from each population. Define $\bar{\alpha}$ in the obvious way and test $H_0 : \alpha_k = \alpha_l$ against $H_1 : \alpha_k \neq \alpha_l$:

Minimize $S = \sum_{j=1}^n (Y_{kj} - \alpha_k - \beta x_{kj})^2 + \sum_{j=1}^n (y_{lj} - \alpha_l - \beta x_{lj})^2$ under H_1, H_0 to get R_1, R_0. We can reason that we should reject H_0 if $R_1 - R_0$ is large; $R_0 = (R_0 - R_1) + R_1$. We find $R_1 \sim \sigma^2\chi_{2n-3}$ whether or not H_0 is true while $R_0 - R_1 \sim \sigma^2\chi_{3-2}^2$ (the degrees of freedom being those of H_1 - those of H_0) if H_0 is true. So we perform a one-tailled test of $T = \frac{\frac{R_0-R_1}{3-2}}{\frac{R_1}{50-3}}$ against the $F_{1,47}$ distribution.

The remainder of the lectures for this course is non-examinable.