IB Statistics (Lent Term, 2025)

Qingyuan Zhao

March 18, 2025

Outline

What is Statistics?

- Inverse probability?
- Data analysis?
- Machine learning/AI?

Focus of this course: statistical inference

- Deductive inference: necessary (e.g. Euclidean geometry)
- Inductive inference: non-necessary (e.g. smoking causes cancer)

The point is we would like to use deduction to justify and guide induction.

Where is Statistics applied?

Statistical inference is needed to answer questions such as:

- What are the voting intentions before an election? (Market research, surveys)
- What is the effect of obesity on life expectancy? (Epidemiology)
- What is the average benefit of a new cancer therapy? (Clinical trials)
- What proportion of temperature change is due to man? (Environmental statistics)
- What is the benefit of speed cameras? (Traffic studies)
- What portfolio maximises expected return? (Financial and actuarial applications)
- How confident are we the Higgs Boson exists? (Physics)
- What are possible benefits and harms of genetically-modified plants? (Agriculture)
- What proportion of the UK economy involves illegal drugs? (Official statistics)
- What is the chance Liverpool will best Arsenal next week? (Sport)

Typical setting: Parametric inference

- Let $X_1, \ldots, X_n$ be independent and identically distributed (IID) random variables taking values in some space $\mathcal{X}$. Let $X = (X_1, \ldots, X_n)^T$.
- We assume the distribution of X_1 belongs to some statistical model $\{p(x; \theta); \theta \in \Theta\}$ but θ is unknown.
  - Example 1: $X_1 \sim \text{Poisson}(\lambda)$, $\theta = \lambda \in \Theta = (0, \infty)$.
  - Example 2: $X_1 \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

Common questions

- Give an estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$ of the true value of θ.
- Give an interval estimator $(\hat{\theta}(X), \hat{\theta}_2(X))$ of θ.
- Testing some hypothesis about θ, e.g. $H_0 : \theta = 0$: is there evidence against H_0?

Big assumption

In general, we need to know the family of distributions involved.

- This can be relaxed for some results (e.g. bias-variance tradeoff).

Review: Probability and random variable

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space.

- Ω is the sample space of outcomes of an experiment.
- $\mathcal{F}$ collects subsets of Ω called events and needs to be a σ-algebra (containing Ω, closed under complementation and countable unions).
- $\mathrm{P} : \mathcal{F} \to [0, 1]$ is the probability measure that satisfies $\mathrm{P}(\Omega) = 1$, $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$, and $\mathrm{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$ for all sequences $(A_i)_{i=1}^{\infty}$ of disjoint events.

A random variable (RV) is a function $X : \Omega \to \mathbb{R}$. *[Example: tossing 2 coins.]*

Review: Probability distribution of a random variable

The cumulative distribution function (CDF) of X is $\mathrm{P}(x) = \mathrm{P}(X \leq x)$.[1]

- Right-continuous and monotone increasing.
- "Uniquely identifies" a RV.

Often, we characterize X via its probability density function (PDF) $\mathrm{p}(x)$.

- When X is discrete (takes values in a countable set $\mathcal{X} \subset \mathbb{R}$), this is usually called the probability mass function $\mathrm{p}(x) = \mathrm{P}(X = x)$.
- When X is continuous, the density function satisfies

$$\mathrm{P}(X \in A) := \mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\}) = \int_A \mathrm{p}(x)dx, \quad \text{for all } A \subseteq \mathbb{R}\,.$$

These definitions naturally extend to random vectors $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$.

[1] The intentional abuse of notation here highlights that we treat RV as an *equivalence class* of functions that have the same probability distribution. In most other texts, CDF is denoted as F and PDF is denoted as f.

Review: Moments

The expectation of X is[2]

$$\mathrm{E}(X) = \begin{cases} \sum_{x \in \mathcal{X}} x\, \mathrm{p}(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x\, \mathrm{p}(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

The variance of X is $\mathrm{Var}(X) = \mathrm{E}\{(X - \mathrm{E}(X))^2\} = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$. The covariance between X and Y is $\mathrm{Cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\}$.

Moment generating function (MGF)

The MGF of X is $M(t) = \mathrm{E}(e^{tX})$, provided that the expectation exists for t in a neighbourhood of 0.

- Relationship with moments: $\mathrm{E}(X^n) = \frac{d^n}{dt^n} M(t)|_{t=0}$.
- Under mild conditions, $M_X = M_Y \Longrightarrow \mathrm{P}_X = \mathrm{P}_Y$.

[2] This is only well-defined if $\mathrm{E}(|X|) < \infty$.

Review: Independence

We say RVs $X_1, X_2, \ldots, X_n$ are independent if

$$\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathrm{P}(X_1 \leq x_1) \ldots \mathrm{P}(X_n \leq x_n), \text{ for all } x_1, \ldots, x_n \in \mathbb{R}.$$

If $X_1, \ldots, X_n$ have probability density/mass functions $\mathrm{p}_{X_1}, \ldots, \mathrm{p}_{X_n}$, then the joint density function of $X = (X_1, \ldots, X_n)$ is given by

$$\mathrm{p}_X(x) = \prod_{i=1}^{n} \mathrm{p}_{X_i}(x_i).$$

- *[Example: Maximum of independent RVs.]*
- *[Example: Distribution of the sum of independent Poisson RVs.]*

Review: Linear transformations

For any RVs $X_1, \ldots, X_n$ and $a_1, \ldots, a_n \in \mathbb{R}$,

$$\mathrm{E}(a_1 X_1 + \ldots a_n X_n) = a_1 \,\mathrm{E}(X_1) + \ldots a_n \,\mathrm{E}(X_n),$$

$$\mathrm{Var}(a_1 X_1 + \ldots a_n X_n) = \sum_{i,j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

If $X_1, \ldots, X_n$ are independent, then $\mathrm{E}(X_1 \cdots X_n) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_n)$, which implies

$$\mathrm{Var}(a_1 X_1 + \ldots a_n X_n) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i).$$

Vector notation

Let $X = (X_1, \ldots, X_n)^T$ be a random vector and $a = (a_1, \ldots, a_n)^T$ be fixed, then

$$\mathrm{E}(a^T X) = a^T \,\mathrm{E}(X) \quad \text{and} \quad \mathrm{Var}(a^T X) = a^T \,\mathrm{Var}(X) a.$$

Review: Change of variables (in 2D)

Consider a differentiable bijection $f : \mathbb{R}^2 \to \mathbb{R}^2$. Then the PDF of $(U, V) = f(X, Y)$ is given by

$$\mathrm{p}_{U,V}(u, v) = \mathrm{p}_{X,Y}(x, y)|\det J(u, v)| = \mathrm{p}_{X,Y}(f^{-1}(u, v), f^{-1}(u, v))|\det J(u, v)|,$$

where $J(u, v)$ is the Jacobian matrix:

$$J(u, v) = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}.$$

Review: Limit theorems

Let $X_1, \ldots, X_n$ be IID RVs with $\mathrm{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$.

- Write the sum as $S_n = \sum_{i=1}^{n} X_i$ and sample mean as $\bar{X}_n = S_n/n$.

Law of large numbers

- Weak law: $\bar{X}_n \xrightarrow{p} \mu$, which means $\mathrm{P}(|\bar{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$.
- Strong law: $\bar{X}_n \xrightarrow{a.s.} \mu$, which means $\mathrm{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$. *[What's the event here?]*

Central limit theorem

$$Z_n = \frac{(S_n - n\mu)}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1),$$

where $\xrightarrow{d}$ means *convergence in distribution*, $\mathrm{P}(Z_n \leq z) \to \Phi(z)$ for all $z \in \mathbb{R}$ and $\Phi(\cdot)$ is the CDF of $\mathrm{N}(0,1)$.

Review: Marginalization and conditioning

Let $\mathrm{p}(x, y)$ denote the probability density/mass function of (X, Y).
The marginal PDF of X is given by

$$\mathrm{p}_X(x) = \begin{cases} \sum_{y \in \mathcal{Y}} f_{X,Y}(x, y), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, & \text{if } Y \text{ is continuous.} \end{cases}$$

The conditional PDF of Y given $X = x$ is given by[3]

$$\mathrm{p}_{Y|X}(y \mid x) = \frac{\mathrm{p}_{X,Y}(x, y)}{\mathrm{p}_X(x)}.$$

These definitions naturally extend to random vectors.

[3]The conditional PDF $\mathrm{p}(y \mid x)$ is not well-defined when $\mathrm{p}_X(x) = 0$, but this does not matter as this "event" has probability 0.

Review: Conditional expectation

The conditional expectation of Y given X is *[Why is this a RV?]*

$$\mathrm{E}(Y \mid X) = \begin{cases} \sum_{y \in \mathcal{Y}} y f(y \mid X), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} y f(y \mid X) dy, & \text{if } Y \text{ is continuous.} \end{cases}$$

Law of total expectation: $\mathrm{E}\{\mathrm{E}(Y \mid X)\} = \mathrm{E}(Y)$.
Law of total variance: $\mathrm{Var}(Y) = \mathrm{E}\{\mathrm{Var}(Y \mid X)\} + \mathrm{Var}\{\mathrm{E}(Y \mid X)\}$.

*Interlude: Hans Rosling

Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC

Important discrete distributions: Binomial

X has the binomial distribution with parameters $n \in \mathbb{N}$ and $0 \leq p \leq 1$, $X \sim \text{Bin}(n, p)$, if

$$\text{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n.$$

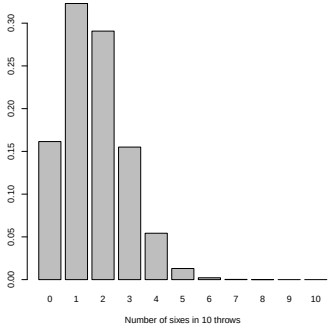
- This is the distribution of the number of successes out of n independent Bernoulli trials, each of which has success probability p.
- $\text{E}(X) = np$, $\text{Var}(X) = np(1-p)$.

Example: Throwing dice

Let X = number of sixes when throw 10 fair dice, so $X \sim \text{Bin}(10, 1/6)$.

[R code]

```
barplot(dbinom(0:10, 10, 1/6), names.arg=0:10,
        xlab="Number of sixes in 10 throws" )
```



Important discrete distributions: Poisson

X has the Poisson distribution with parameter $\lambda > 0$, $X \sim \text{Poisson}(\lambda)$, if

$$\text{P}(X = x) = e^{-\lambda}\lambda^x/x!, \ x = 0, 1, 2, \ldots.$$

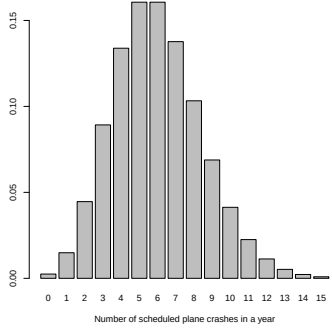
- $\text{E}(X) = \lambda$ and $\text{Var}(X) = \lambda$.
- The Poisson(λ) is the limit of Bin(n, p) when $n \to \infty$ and $np \to \lambda$. *[ES1]*
- *In a Poisson process, the number of events $X(t)$ in an interval of length t is Poisson(λt), where λ is the intensity of the process (rate per unit time).

Example: Plane crashes

Assume scheduled plane crashes occur as a Poisson process with a rate of 1 every 2 months. How many (X) will occur in a year?

Number in two months is Poisson(1), and so $X \sim$ Poisson(6).

```
barplot(dpois(0:15, 6), names.arg=0:15,
        xlab="Number of scheduled plane crashes in a year" )
```



Important discrete distributions: Negative Binomial

X has the negative binomial distribution with parameters $k \in \mathbb{N}$ and $0 < p < 1$, $X \sim \text{NegBin}(k, p)$, if

$$\text{P}(X = x) = \binom{x-1}{k-1}(1-p)^{x-k}p^k, \; x = k, k+1, \ldots.$$

- This is the distribution of the number of trials up to the k-th success, in a sequence of independent Bernoulli trials each with success probability p.
- $\text{E}(X) = k/p$, $\text{Var}(X) = k(1-p)/p^2$.
- When $k = 1$, this is called the geometric distribution with parameter p.
- Some texts use negative binomial to refer to the distribution of $Y = X - k$ (number of failures before the k-th success). Be careful!

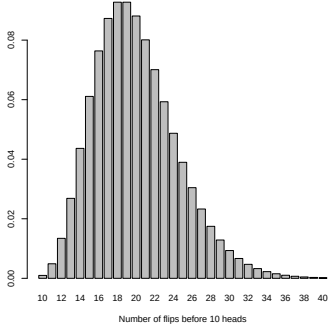
Example: Coin flip

How many times do I have to flip a coin before I get 10 heads?

This is the first (X) definition of Negative Binomial since it includes all the flips.

R uses second definition (Y) of Negative Binomial, so need to add in the 10 heads:

```
barplot(dnbinom(0:30, 10, 1/2), names.arg=0:30 + 10,
        xlab="Number of flips before 10 heads")
```



Important discrete distributions: Multinomial

Suppose we have a sequence of n independent trials where at each trial there are $k \in \mathbb{N}$ possible outcomes, and that at each trial the probability of outcome i is p_i. Let N_i be the number of times outcome i occurs.
We say $(N_1, \ldots, N_k)$ follows the multinomial distribution with parameters n and $p = (p_1, \ldots, p_k)$, with joint PMF

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \begin{cases} \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, & \text{if } \sum_{i=1}^k n_i = n, \\ 0, & \text{otherwise.} \end{cases}$$

- $n \in \mathbb{N}$ and p in the standard simplex in $\mathbb{R}^k$: $p_i \geq 0$ for all i and $\sum_{i=1}^k p_i = 1$.
- The RVs $N_1, \ldots, N_k$ are not independent. *[Why?]*
- The marginal distribution of N_i is $\mathrm{Bin}(n, p_i)$.
- Example: I throw 6 dice. The probability I get one of each face is $6!/6^6 \approx 0.015$. (R code is `dmultinom(x=rep(1,6), size=6, prob=rep(1/6,6))`.)

Important continuous distributions: Uniform

X has the uniform distribution on $[a, b]$, $X \sim \text{Unif}[a, b]$ $(-\infty < a < b < \infty)$, if it has PDF

$$\text{p}(x) = \frac{1}{b-a}, \quad x \in [a, b].$$

- $\text{E}(X) = (a+b)/2$ and $\text{Var}(X) = (b-a)^2/12$.

Important continuous distributions: Gamma

X has the Gamma distribution with shape $\alpha > 0$ and rate $\lambda > 0$, $X \sim \text{Gamma}(\alpha, \lambda)$, if it has PDF

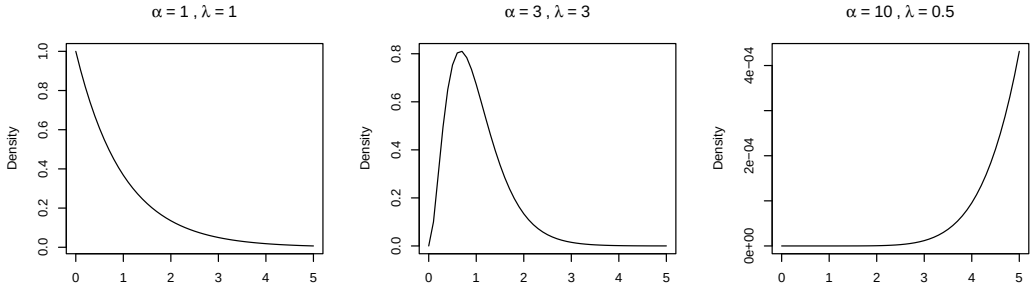
$$p(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0,$$

where $\Gamma(\alpha)$ is the gamma function defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$.

- $E(X) = \alpha/\lambda$ and $\text{Var}(X) = \alpha/\lambda^2$.
- If $\alpha = 1$, this is the exponential distribution: $\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$.
- If $X_i \sim \text{Gamma}(\alpha_i, \lambda), i = 1, \ldots, n$ are independent, then $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \lambda)$. *[Prove this via MGF.]*
- If $X \sim \text{Gamma}(\alpha, \lambda)$, then $cX \sim \text{Gamma}(\alpha, \lambda/c)$ for any $c > 0$.
- Note the following results for the gamma function:
  1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$;
  2. if $n \in \mathbb{N}$ then $\Gamma(n) = (n-1)!$.

Density plot of some Gamma distributions

```
par(mfrow = c(1, 3))
alpha<-c(1, 3, 10); lambda<-c(1, 3, 0.5); x<-seq(0, 5, 0.1)
for(i in 1:3) {
  y= dgamma(x, alpha[i], lambda[i])
  plot(x, y, type = "l", ylab = "Density",
       main = bquote(alpha~"="~.(alpha[i])~","~lambda~"="~.(lambda[i])))
}
```



Important continuous distributions: Beta

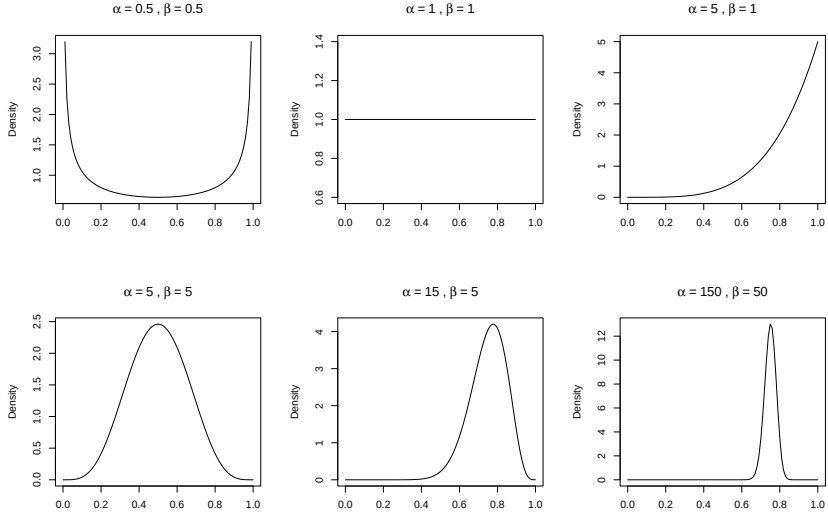
Suppose $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$ are independent.
The RV $Z = X/(X+Y)$ follows the Beta distribution with parameters $\alpha, \beta > 0$ and has the PDF *[ES1]*

$$\mathrm{p}(z) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < x < 1,$$

where $B(\alpha, \beta)$ is the Beta function defined by $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

- $\mathrm{E}(Z) = \alpha/(\alpha+\beta)$ and $\mathrm{Var}(Z) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- The mode is $(\alpha-1)/(\alpha+\beta-2)$.
- Note that $\mathrm{Beta}(1,1) = \mathrm{Unif}[0,1]$.

Density plot of some Beta distributions



Important continuous distributions: Normal

X has the normal (or Gaussian) distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, $X \sim \mathrm{N}(\mu, \sigma^2)$, if it has PDF

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

- $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.
- If $\mu = 0$ and $\sigma^2 = 1$, this is the standard normal distribution. We use ϕ to denote its PDF and Φ for its CDF. *[Write down $\phi(x)$ and $\Phi(x)$.]*
- The $(1-\alpha)$ quantile (upper $100\alpha\%$ percentile) of $\mathrm{N}(0,1)$ is denoted as z_α, so

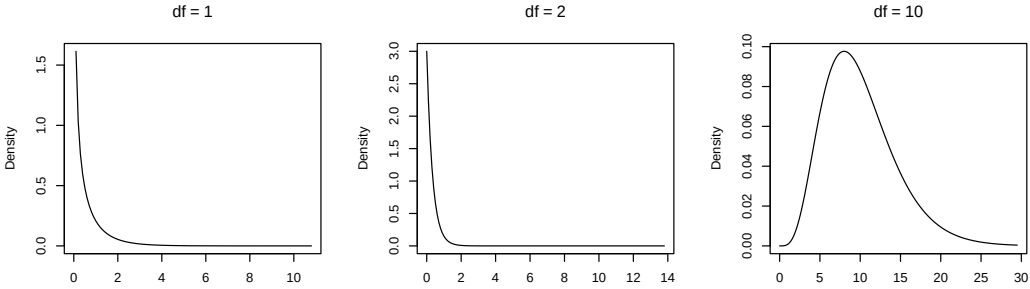
$$\mathrm{P}(Z > z_\alpha) = \alpha, \text{ where } Z \sim \mathrm{N}(0,1).$$

```
> qnorm(c(0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99))
[1] -2.3263479 -1.6448536 -0.6744898  0.0000000  0.6744898  1.6448536  2.3263479
```

Important continuous distributions: Chi-squared

If $Z_1, \ldots, Z_k \overset{\text{IID}}{\sim} \text{N}(0,1)$, then $X = \sum_{i=1}^{k} Z_i^2$ has the chi-squared distribution with k degrees of freedom, $X \sim \chi^2_k$.

- Since $\text{E}(Z_i^2) = 1$ and $\text{E}(Z_i^4) = 3$, we find that $\text{E}(X) = k$ and $\text{Var}(X) = 2k$.
- $\chi^2_k = \text{Gamma}(k/2, 1/2)$. *[Prove this via MGF.]*
- We denote its $(1-\alpha)$ quantile by $\chi^2_k(\alpha)$, so if $X \sim \chi^2_k$ then $\text{P}(X > \chi^2_k(\alpha)) = \alpha$.



Outline

Point estimation

Suppose we observe data X that follows a probability distribution with PDF $\mathrm{p}(x;\theta)$ and $\theta \in \Theta$ is unknown.

- Often, $X = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are IID.

Definition

An estimator is a statistic (meaning a function of the data) $\hat{\theta} := T(X)$ which we use to approximate the true parameter θ.

- If we observe $X = x = (x_1, \ldots, x_n)$, then our estimate is $T(x)$.
- The distribution of $T(X)$ is called its sampling distribution.

Example

Suppose $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, 1)$.

- A possible estimator for μ is the sample mean $\hat{\mu} = T(X) = \sum_{i=1}^{n} X_i/n$.
- The sampling distribution is $\hat{\mu} \sim \mathrm{N}(\mu, 1/n)$.

Bias

Definition

The bias of $\hat{\theta} = T(X)$ for θ is defined as

$$\text{bias}(\hat{\theta}) = \mathsf{E}_\theta(\hat{\theta}) - \theta.$$

We say $\hat{\theta}$ is unbiased if $\text{bias}(\hat{\theta}) = 0$ for all $\theta \in \Theta$.

- Bias is generally a function of θ, which is not explicit in our notation.
- This definition naturally extends to estimators of a given transformation of θ.

Example (cont'd)

- The sample mean $\hat{\mu} = \sum_{i=1}^{n} X_i/n$ is unbiased for μ. *[Why?]*
- The sample median $X_{(\frac{n+1}{2})}$ (assuming n is odd) is also unbiased for μ. *[Why? What about even n?]*

Mean squared error

Definition

The mean squared error (MSE) of an estimator $\hat{\theta}$ is defined as

$$\text{MSE}(\hat{\theta}) = \text{E}_\theta\{(\hat{\theta} - \theta)^2\}.$$

- Like the bias, MSE is a function of θ.

Proposition: Bias-variance decomposition

$$\text{MSE}(\hat{\theta}) = \text{Var}_\theta(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

- *[Prove this.]*
- There is often a tradeoff between bias and variance. Increasing bias slightly may lead to greater reduction in variance and overall smaller MSE.

*Interlude: Child bed fever mortality

- The surprising history of hand-washing (BBC REEL).
- Analysis of child bed fever mortality by Ignaz Semmelweis.

Example: Estimators for Binomial mean

Suppose $X \sim \text{Bin}(n, \theta)$, n is known.

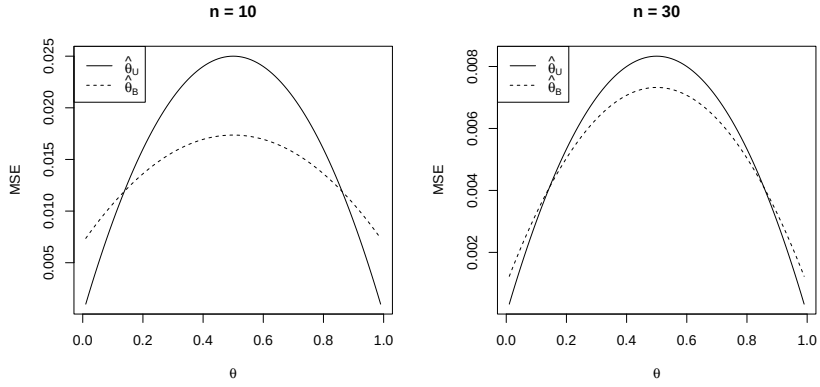
1. The standard estimator of θ is $\hat{\theta}_U = X/n$.
   - $\text{bias}(\hat{\theta}_U) = 0$ and $\text{MSE}(\hat{\theta}_U) = \theta(1-\theta)/n$. *[Why?]*
2. An alternative estimator is

$$\hat{\theta}_B = \frac{X+1}{n+2} = w\hat{\theta}_U + (1-w)\frac{1}{2}, \quad \text{where} \quad w = \frac{n}{n+2}.$$

   - When $X = 8$ and $n = 10$, we have $\hat{\theta}_U = 0.8$ and $\hat{\theta}_B = 0.75$.
   - $\text{bias}(\hat{\theta}_B) = (n\theta + 1)/(n+2) = (1-w)(1/2 - \theta)$.
   - $\text{MSE}(\hat{\theta}_B) = w^2\theta(1-\theta)/n + (1-w)^2(1/2-\theta)^2$.

Which estimator is better?

Example: Estimators for Binomial mean (cont'd)



- So the biased estimator has smaller MSE in much of the range of θ.
- Whether $\hat{\theta}_B$ is preferable depends on our prior judgement about θ.

Why unbiasedness is often too strong

Suppose $X \sim \text{Poisson}(\lambda)$ and we'd like to estimate $\theta = \{\text{P}(X = 0)\}^2 = e^{-2\lambda}$.

- The only unbiased estimator of θ is $\hat{\theta} = T(X) = (-1)^X$. *[Prove this.]*
- But $\hat{\theta}$ is not a sensible estimator.

*Decision theory

To guess is cheap, to guess wrongly is expensive. (Old Chinese proverb)

- A decision rule $d : \mathcal{X} \to \mathcal{D}$ maps data to decision. *[What's $\mathcal{D}$ in point estimation?]*
- A basic notion in statistical decision theory is the risk function:

$$R(\theta, d) = \mathrm{E}\{L(\theta, d(X))\},$$

  where $L : \Theta \times \mathcal{D} \to \mathbb{R}$ is called the loss function (economists call $-L$ the utility).
- Further reading: expected utility hypothesis.

Natural question: optimality in a class of decision rules

- Example: uniform minimum variance unbiased estimator (UMVUE). *[ES1]*
- Example: $\hat{\theta}$ is called minimax optimal if it solves $\min_{\hat{\theta}} \max_{\theta \in \Theta} R(\theta, \hat{\theta})$.
- Often, it is challenging to even "ask the right question".

Outline

Sufficient statistics

Question: Is there a statistic that contains all the useful information in the data?

Example: Bernoulli trials

$X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\theta)$, so

$$p(x;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{\sum x_i}.$$

This depends on x only through $T(x) = \sum_{i=1}^{n} x_i$.

Definition

A statistics $T = T(X)$ is sufficient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Factorization criterion

Suppose the PDF of X is $\mathrm{p}(x;\theta)$.

Theorem

T is sufficient (for θ) iff $\mathrm{p}(x;\theta) = g(T(x),\theta)h(x)$ for some suitable functions g and h.

- *[Prove this (in discrete case).]*

Example: Bernoulli trials (cont'd)

$T(X) = \sum_{i=1}^{n} X_i$ is sufficient because $\mathrm{p}(x;\theta) = (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{\sum x_i}$. *[Why?]*

Example: Uniform RVs

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{Unif}[0,\theta]$, then $T(X) = \max_i X_i$ is sufficient. *[Prove this.]*

Minimal sufficiency

Sufficient statistics are not unique.

- The order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ ($X_{(k)}$ is the kth smallest value in $X_1, \ldots, X_n$) are always sufficient. *[Why?]*

Definition

A sufficient statistic $T(X)$ is minimal sufficient if it is a function of every other sufficient statistic: if $S(X)$ is also sufficient, $S(x) = S(y) \Rightarrow T(x) = T(y)$ for all x, y.

Theorem

Suppose $T(x) = T(y)$ if and only if $\mathrm{p}(x;\theta)/\mathrm{p}(y;\theta)$ does not depend on θ (if the range of X depends on θ, this means $\mathrm{p}(x;\theta) = c(x,y)\,\mathrm{p}(y;\theta)$ for some function $c(x,y)$). Then T is minimal sufficient.

- *[*Proof sketch.]*
- *[Example: If $X_1, \ldots, X_n \overset{IID}{\sim} N(\mu, \sigma^2)$, show that $(\bar{X}, S^2 := \sum(X_i - \bar{X})^2/(n-1))$ is minimal sufficient.]*

*Interlude: John Snow and cholera

Continuing interlude 1: Semmelweis's views were much more favorably received in the UK than on the continent, but he was more often cited than understood (Wikipedia).

- Semmelweis "made out very conclusively" that "miasms derived from the dissecting room will excite puerperal disease." (W. Tyler Smith, 1856)

Now watch John Snow and the 1854 Broad Street cholera outbreak (Harvard Online).

Snow's statistical analysis (most striking table, full report over 200 pages)

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

- S&V took water from a heavily contaminated stretch of the Thames
- Lambeth moved its intake pipe upstream in 1852 to get purer water.

Why is sufficiency useful?

The sufficiency principle says that if $T(X)$ is a sufficient statistic, and if in two experiments with data x and y we have $T(x) = T(y)$, then the "evidence" about θ given by the two experiments is the same.

- This is not a mathematical theorem but a "principle".

Theorem (Rao-Blackwell)

Suppose $\Theta \in \mathbb{R}$ and T is a sufficient statistic. For any estimator $\tilde{\theta}$ of θ, we have

$$\mathrm{MSE}(\tilde{\theta}) \geq \mathrm{MSE}(\hat{\theta}),$$

where $\hat{\theta} = \mathrm{E}_\theta(\tilde{\theta} \mid T)$. The equality holds if and only if $\tilde{\theta}$ is a function of T.

- *[Prove this. Why is $\hat{\theta}$ an estimator (doesn't depend on θ)?]*
- This gives us a strong reason to only consider estimators that are functions of sufficient statistics.

Examples of Rao-Blackwellization

1. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$ and let $\theta = e^{-\lambda}$.
   - The minimal sufficient statistic is $T = \sum_{i=1}^{n} X_i$. *[Why?]*
   - An easy unbiased estimator of θ is $\tilde{\theta} = 1_{\{X_1=0\}}$, but a much better estimator is
   $$\hat{\theta} = \mathrm{E}_\theta(\tilde{\theta} \mid T) = \left(1 - \frac{1}{n}\right)^T. \quad \textit{[Derive this.]}$$
2. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Unif}[0, \theta]$.
   - We have shown that $T = \max_{1 \leq i \leq n} X_i$ is sufficient.
   - *[What is an unbiased estimator of θ using just X_1?]*
   - *[Derive the corresponding Rao-Blackwell estimator.]*

Outline

Likelihood

Definition

Suppose X has joint PDF $\mathrm{p}(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$.

- The likelihood of θ is the density at the observed data, viewed as a function of θ, i.e. $L : \theta \mapsto \mathrm{p}(X;\theta)$.
- The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes θ, i.e. $\hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta L(\theta)$.

IID setting

- If $X = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are IID, each with PDF $\mathrm{p}(x;\theta)$, then

$$L(\theta) = \prod_{i=1}^{n} \mathrm{p}(X_i;\theta).$$

- It is often easier to maximize the log-likelihood: $l(\theta) = \sum_{i=1}^{n} \log \mathrm{p}(X_i;\theta)$.

Examples

1. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$. The MLE of p is $\hat{p} = \sum_{i=1}^n X_i/n$.
2. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{N}(\mu, \sigma^2)$. The MLE is

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{XX}/n), \text{ where } S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2.$$

3. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Unif}[0, \theta]$. The likelihood function is given by

$$L(\theta) = \frac{1}{\theta^n} 1_{\{\max_i X_i \leq \theta\}}.$$

So the MLE is $\hat{\theta} = \max_i X_i$. *[Sketch the likelihood function.]*

[Derive the MLEs. Which are (asymptotically) unbiased?]

Properties of MLEs

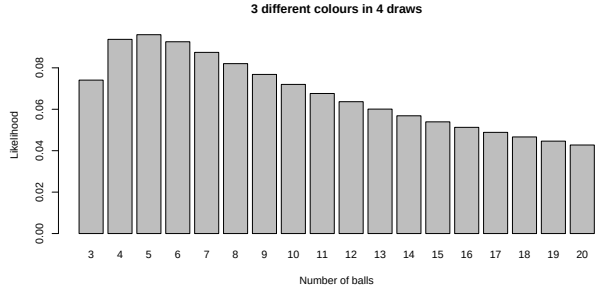
1. The MLE is a function of any sufficient statistic. *[Why?]*
2. If $\eta = h(\theta)$ and h is a bijection, then the MLE of η is $\hat{\eta} = h(\hat{\theta})$.
3. *In "regular" IID settings, the MLE is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, I(\theta)^{-1}), \text{ as } n \to \infty,$$

   where $I(\theta)$ is known as the Fisher information matrix and $I(\theta)^{-1}$ is the "smallest possible" variance. *[Part II PoS.]*
4. Often the MLE has no closed form and needs to be found numerically.

Further example

Suppose a basket contains k balls of different colours, and k is unknown. We are allowed to sample with replacement, the first four balls are Red, Purple, Red, Yellow. The likelihood for k is $L(k) = \frac{(k-1)(k-2)}{k^3}$. *[Derive this.]*



- MLE is $\hat{k} = 5$, but likelihood is fairly flat.

*Interlude: How many words did Shakespeare know?

Shakespeare's known works comprise $n = 884,647$ words. He wrote 31,534 different words, of which 14,376 appear only once, 4,343 twice, etc. Suppose he knew s words.

- Suppose $X_i(t) =$ no. times word i in a sample of nt words $\sim$ Poisson$(\lambda_i t)$.
- Let N_j be the number of words which occur j times in a sample of n words.
- Let $M(t)$ be the number of distinct words in a sample of $n(1+t)$ words that does not appear among the first n words.
- It is can be shown that

$$\mathrm{E}(N_j) = \sum_{i=1}^{s} e^{-\lambda_i} \frac{\lambda_i^j}{j!}, \quad \mathrm{E}(M(t)) = \sum_{i=1}^{s} e^{-\lambda_i}(1 - e^{-\lambda_i t}) = \sum_{j=1}^{\infty} (-1)^{j-1} t^j \, \mathrm{E}(N_j).$$

- By replacing $\mathrm{E}(N_j)$ with N_j, we obtain an unbiased estimator of $\mathrm{E}(M(t))$. For Shakespeare's data and $t = 1$, $\widehat{\mathrm{E}(M(1))} = 11,430$.
- $\widehat{\mathrm{E}(M(\infty))}$ does not converge, but a more complicated method suggests that Shakepeare knew at least 35,000 more words (Efron and Thisted, 1976).

*A bit of history

Method of moment (MoM) estimator

Idea: Match sample moments with theoretical moments.

- First introduced by Chebyshev (1887) in probability to prove CLT.
- Use in statistics dates back at least to Karl Pearson (early 1900s).
- Simple but usually less efficient than MLE.

R A Fisher (1921). "On the Mathematical Foundations of Theoretical Statistics"

- First appearance of many fundamental concepts: consistency, efficiency, estimation, likelihood, sufficiency, parameter. (Further reading: Fisher in 1921 by Stigler).
- The likelihood principle says that all the information about θ obtainable from an experiment is contained in the likelihood function for θ given X. This is stronger than the sufficiency principle.

Outline

Confidence intervals

A distinguishing feature of statistical inference (compared to ML/AI) is its emphasis on uncertainty quantification. Let $\alpha \in [0,1]$ be given and suppose $\theta \in \Theta \subseteq \mathbb{R}$.

Definition

A $(1-\alpha)$-confidence interval (CI) for θ is an interval $[L(X), U(X)] \subseteq \Theta$ such that

$$\mathrm{P}_\theta(L(X) \leq \theta \leq U(X)) = 1-\alpha, \text{ for all } \theta \in \Theta.$$

- In this statement, the CI is random and θ is fixed.
- Correct frequentist interpretation: if we repeat the experiment many times, on average $100(1-\alpha)\%$ of the time, the interval $[L(X), U(X)]$ covers θ.
- Wrong interpretation: having observed $X = x$, the interval $[L(x), U(x)]$ contains θ with probability $1-\alpha$. *[What is this probability?]*
- This definition can be naturally extended to vector-valued θ. *[How?]*
- Often difficult to find exact CIs. May be enough to have $\mathrm{P}(\text{cover}) \geq$ or $\approx 1-\alpha$.

Example: Normal location problem

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\theta, 1)$. The following is a $(1-\alpha)$-CI for θ:

$$[\bar{X} - z_{\alpha_1}/\sqrt{n}, \bar{X} + z_{\alpha_2}/\sqrt{n}], \text{ whenever } \alpha_1, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = \alpha.$$

- *[Derive this.]* (Recall that z_α means the upper-α quantile of $\mathrm{N}(0,1)$.)
- Typically, we try to centre the CI around the point estimator and minimize the length. So it is sensible to choose $\alpha_1 = \alpha_2 = \alpha/2$.
- When $\alpha = 0.05$, $z_{\alpha/2} \approx 1.96$.

General recipe to construct CIs

1. Find a pivotal quantity $R(X, \theta)$ such that the distribution of $R(X, \theta)$ under P_θ does not depend on θ.
2. By using appropriate quantiles of $R(X, \theta)$, find c_1, c_2 such that

$$\mathrm{P}(c_1 \leq R(X, \theta) \leq c_2) = 1 - \alpha$$

3. Rearrange the inequalities to leave θ in the middle.

Proposition

If $[L(X), U(X)]$ is a $(1-\alpha)$-CI of θ and $h : \Theta \to \mathbb{R}$ is monotone increasing, then $[h(L(X)), h(U(X))]$ is a $(1-\alpha)$-CI of $h(\theta)$.

*Interlude: bias in polling

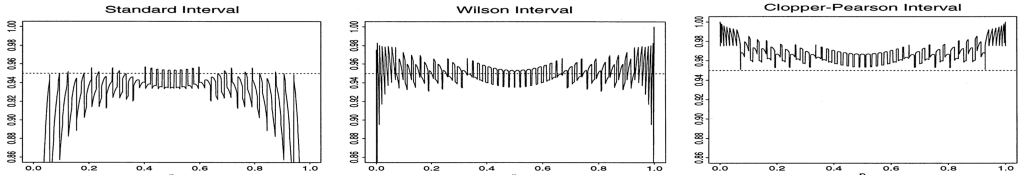
- What polls can actually tell us (Vox).
  - Interesting statistical concepts: "margin of error", "weighting", "population".
- Further readings:
  1. The model exactly predicted the most likely election map (Nate Silver).
  2. Disentangling Bias and Variance in Election Polls (Shirani-Mehr et al.).

We find that average survey error as measured by root mean square error is approximately 3.5 percentage points, about twice as large as that implied by most reported margins of error. We decompose survey error into election-level bias and variance terms. We find that average absolute election-level bias is about 2 percentage points, indicating that polls for a given election often share a common component of error.

Example: Binomial proportion

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$.

- An approximate $(1-\alpha)$-CI for p is $\left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ where $\hat{p} = \sum_i X_i / n$.
- Wilson: use the pivot $(\hat{p} - p)/\sqrt{p(1-p)/n}$ instead.
- Clopper-Pearson ("exact"): use the binomial distribution directly.



Coverage probability for $n = 50$ (Brown, Cai, DasGupta, 2001).

*Example: Binomial proportion (numerical)

Output of R code:

```
> t(sapply(c("wald", "wilson", "clopper-pearson"),
+          function (method) {DescTools::BinomCI(200, 1000, method = method)}))
                  [,1]      [,2]      [,3]
wald               0.2 0.1752082 0.2247918
wilson             0.2 0.1763771 0.2259190
clopper-pearson    0.2 0.1756206 0.2261594
> t(sapply(c("wald", "wilson", "clopper-pearson"),
+          function (method) {DescTools::BinomCI(50, 1000, method = method)}))
                  [,1]       [,2]       [,3]
wald              0.05 0.03649188 0.06350812
wilson            0.05 0.03813026 0.06531382
clopper-pearson   0.05 0.03733540 0.06539049
> t(sapply(c("wald", "wilson", "clopper-pearson"),
+          function (method) {DescTools::BinomCI(5, 100, method = method)}))
                  [,1]        [,2]       [,3]
wald              0.05 0.007283575 0.09271642
wilson            0.05 0.021543679 0.11175047
clopper-pearson   0.05 0.016431879 0.11283491
```

Outline

Frequentist vs. Bayesian

In Probability and Philosophy

- Frequentists interpret probability as long-term frequency ("physical").
- Bayesians interpret probability as subjective plausibility ("evidential").

In Statistics

- Frequentists treat θ as fixed quantities describing a natural law.
- Bayesians treat θ as random variables and view statistical inference as updating one's belief about θ.

The Bayesian paradigm

- Let $\pi(\theta)$ be the PDF of prior distribution on Θ (of the investigator).
- As before, $X \mid \theta \sim \mathrm{p}(x \mid \theta)$. (We use $\mid$ to emphasize θ is conditioned on.)
- By Bayes' theorem, the posterior distribution of θ given $X = x$ has PDF/PMF

$$\pi(\theta \mid x) = \frac{\pi(\theta)\,\mathrm{p}(x \mid \theta)}{\mathrm{p}(x)},$$

  where $\mathrm{p}(x) = \int \mathrm{p}(x \mid \theta)\pi(\theta)d\theta$ is the marginal density of X.
- Often we ignore the normalizing constant and write $\pi(\theta \mid x) \propto \mathrm{p}(x \mid \theta)\pi(\theta)$, i.e.

$$\text{posterior} \propto \text{prior} \cdot \text{likelihood},$$

  so we use information in the likelihood to update our belief about θ.
- Thus, Bayesian inference obeys the likelihood principle and hence the sufficiency principle: $\pi(\theta \mid x)$ depends on x only through sufficient statistics. *[Show this.]*

Example: COVID testing

A patient walked into a COVID testing centre in central Cambridge and had a positive test result. What is the probability that they were actually infected?

- Let $\theta = 1_{\{\text{patient is infected}\}}$ and $X = 1_{\{\text{test is positive}\}}$. We know the test has sensitivity $\mathrm{P}(X = 1 \mid \theta = 1) = 98\%$ and specificity $\mathrm{P}(X = 0 \mid \theta = 0) = 99\%$.
- To choose a prior, we could set $\pi(\theta = 1)$ to the proportion of people infected with COVID-19 in the country at the time.
- Say $\pi(\theta = 1) = 2\%$ using ONS survey, then $\pi(\theta = 1 \mid X = 1) = 2/3$. *[Why?]*
- *[But is this the "right" prior?]*

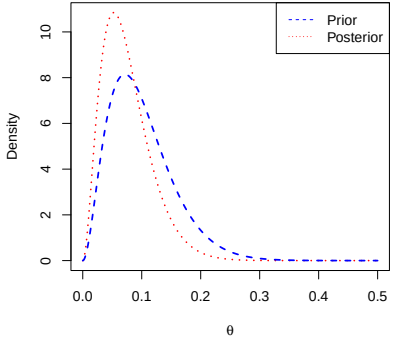
Example: Beta-Binomial

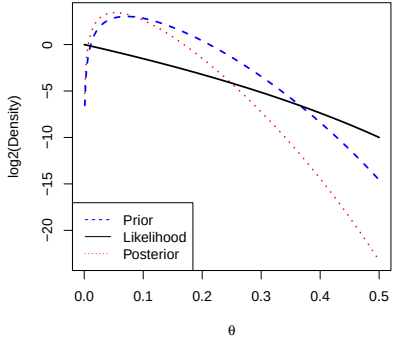
- Prior: $\theta \sim \text{Beta}(\alpha, \beta)$.
- Likelihood model: $X \mid \theta \sim \text{Binom}(n, \theta)$.
- Posterior is $\theta \mid X \sim \text{Beta}(\alpha + X, \beta + (n - X))$. *[Show this.]*
- When $\alpha = \beta = 1$ ("flat prior"), the posterior mean is $(X + 1)/(n + 2)$ ("Laplace estimator" from Lecture 2).

Example: Beta-Binomial (numerical)

θ = mortality rate of a new surgery in Addenbrookes Hospital. No deaths in the first 10 surgeries. Elsewhere, the mortality rate is between 3% and 20% (average 10%).

- Choose $\alpha = 3$ and $\beta = 27$, so $\pi(\theta)$ has mean 0.1 and $\pi(0.03 \leq \theta \leq 0.2) \approx 0.9$.
- The posterior is Beta$(3, 37)$.





*Interlude: Ellsberg Paradox

A bag contains 300 balls, of which 100 are red and 200 are either blue and green. One is drawn at random.

Question 1

Which of the following gambles do you find most attractive?

- You get £1000 if the ball is red (A).
- You get £1000 if the ball is blue (B).

Question 2

Which of the following gambles do you find most attractive?

- You get £1000 if the ball is not red (C).
- You get £1000 if the ball is not blue (D).

Bayesian approach to point estimation

Idea: Use the "centre" of $\pi(\theta \mid X)$ (e.g. mean, median, mode) as an estimator of θ.

Definition

- Let $L : \Theta \times \Theta \to \mathbb{R}$ be a loss function.
- The Bayes risk of an estimator $\tilde{\theta} = \tilde{\theta}(X)$ is defined as

$$R(\tilde{\theta}) = \mathrm{E}_{\pi}(L(\theta, \tilde{\theta}) \mid X) = \int_{\Theta} L(\theta, \tilde{\theta}(X))\pi(\theta \mid X)d\theta.$$

- The Bayes estimator is $\hat{\theta} = \arg\min_{\tilde{\theta}} R(\tilde{\theta})$.

Examples

1. The Bayes estimator is the posterior mean with quadratic loss: $L(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$.
2. The Bayes estimator is the posterior median with absolute loss: $L(\theta, \tilde{\theta}) = |\theta - \tilde{\theta}|$.

[Show these.]

Bayesian approach to interval estimation

Definition

We say $[L(X), U(X)]$ is a $(1-\alpha)$-credible interval for θ if

$$\pi(L(X) \leq \theta \leq U(X) \mid X) = 1 - \alpha.$$

So $[L(X), U(X)]$ has $(1-\alpha)$ posterior probability and can be easily found using quantiles of the posterior distribution. *[Compare to confidence intervals.]*

Example: Beta-Binomial (Prior=Beta(3, 27), Posterior=Beta(3, 37))

	Mean	Median	Mode	Q2.5%	Q97.5%
Prior	0.1	0.091	0.071	0.029	0.20
Posterior	0.075	0.068	0.053	0.021	0.15

- MLE $= 0$.
- Wald CI: [0, 0]; Wilson CI: [0, 0.28]; Clopper-Pearson CI: [0, 0.31] $(1 - \alpha = 95\%)$.

Example: Normal location problem

- Prior: $\mu \sim \mathrm{N}(\mu_0, n_0^{-1})$, $\mu_0 \in \mathbb{R}$, $n_0 > 0$.
- Likelihood model: $X_1, \ldots, X_n \mid \mu \overset{IID}{\sim} \mathrm{N}(\mu, 1)$.
- Posterior:

$$\mu \mid X \sim \mathrm{N}\left(\frac{n_0\mu_0 + n\bar{X}}{n_0 + n}, \frac{1}{n_0 + n}\right). \quad \textit{[Show this.]}$$

- Interpretation: prior is like n_0 "observations" with sample mean μ_0.

[Compare Bayesian point estimator with the MLE.]

Conjugacy

- We say a family of prior distributions is conjugate to a likelihood model if the posterior distribution is in the same family.
- Examples: Beta-Binomial, Normal-Normal (many more on Wikipedia).

*History and remarks

- Bayesian statistics goes back to Bayes and Laplace in late 18th century.
- Frequentist thinking dominated 20th century after Fisher, Neyman, E Pearson.
- Bayesians have made a come-back with more powerful computers and MCMC.

Remarks

- The practical difference is often not big: Berstein-von Mises theorem. *[Part II PoS]*
- The debate has generated better understanding and methodology for both parties.

Score sheet

- Bayesian: 1. Belief (prior); 2. Principled; 3. One distribution; 4. Dynamic; 5. Individual (subjective); 6. Aggressive.
- Frequentist: 1. Behaviour (method); 2. Opportunistic; 3. Many distributions (bootstrap?); 4. Static; 5. Community (objective); 6. Defensive.

From "A 250-Year Argument: Belief, Behavior, and the Bootstrap" by Efron (2012).

Outline

Motivation

- A hypothesis is an assumption about the distribution of the data X.

Examples

1. Is this coin fair?
2. Is the phenotypic ratio exactly 9:3:3:1 for a dihybrid cross?
3. Do students in different Colleges come from the same socioeconomic background?

[How can we formulate these mathematically?]

Two different formulations

1. Fisher: Do the data provide evidence against a null hypothesis?
2. Neyman-Pearson: Having seen the data, shall we choose the null hypothesis or the alternative hypothesis?

My take: N-P is mathematically superior but is prone to mis-interpretation.

Hypothesis testing

Let a statistical model $X \sim \mathrm{p}(x;\theta)$, $\theta \in \Theta$ be given.

Definition

- The null hypothesis is $H_0 : \theta \in \Theta_0$ and the alternative hypothesis is $H_1 : \theta \in \Theta_1$ for some disjoint $\Theta_0, \Theta_1 \subset \Theta$.
- We say H_0 is simple if Θ_0 is a singleton; otherwise H_0 is composite. Same for H_1.
- A test is a binary-valued statistic $T(X) \in \{0, 1\}$.
  - $T(X) = 1$ means "the data contains enough evidence against H_0";
  - $T(X) = 0$ means "the data contains not enough evidence against H_0".
- Avoid saying "H_0 is true/false".

  In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "A is true" and "A is known to be true". —R A Fisher (Barnard, 1992)

The decision-theoretic approach (Neyman-Pearson)

Definition

- The power function of the test is $\beta(T;\theta) = \mathrm{P}_\theta(T(X) = 1)$.
- Type I error rate: $\beta(T;\theta)$ for $\theta \in \Theta_0$ (false "rejection"). *[What is the loss?]*
- Type II error rate: $1 - \beta(T;\theta)$ for $\theta \in \Theta_1$ (false "acceptance").
- The size of the test is defined as $\sup_{\theta\in\Theta_0} \beta(T;\theta)$.

What is the "optimal" test?

- Not an easy question: there is a tradeoff between type I and II errors.
- The Neyman-Pearson theory treats these risks in an asymmetrical way and asks:

How small can the type II error be if the type I error is no large than a given value?

Likelihood ratio tests: simple vs. simple

Suppose H_0 and H_1 are simple, so $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.

- The likelihood ratio statistic is defined as

$$\Lambda(X) = \frac{L(\theta_1)}{L(\theta_0)} = \frac{\mathrm{p}(X;\theta_1)}{\mathrm{p}(X;\theta_0)}.$$

- A likelihood ratio test is defined as $T(X) = 1_{\{\Lambda > c\}}$ (reject H_0 if $\Lambda(X) > c$).

Neyman-Pearson Lemma

Suppose $\mathrm{p}(x;\theta_0)$ and $\mathrm{p}(x;\theta_1)$ are nonzero on the same set. Consider any $\alpha \in (0,1)$. Suppose there exists $c > 0$ such that $T^*(X) = 1_{\{\Lambda(X) > c\}}$ has exactly size α. Then $T^*(X)$ solves

$$\text{maximize} \quad \beta(T;\theta_1) \quad \text{subject to} \quad \beta(T;\theta_0) \leq \alpha.$$

- *[Proof.]*

Remarks on Neyman-Pearson

1. The N-P Lemma says the likelihood ratio test is the most powerful for simple vs. simple. Another tangible form of the likelihood principle.
2. Nice mathematical result: closed-form solution to a linear program over functions.
3. Continuous density ensures a test of exactly size α exists.
4. Why is this not a Theorem? Composite hypotheses in next lecture.

*Interlude: A love story

"You haven't told me yet," said Lady Nuttal, "what your fiancé does for a living?"

"He's a statistician," replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

"But Aunt Sara, it's a very interesting profession," said Lamia warmly.

"I don't doubt it," said her aunt, who obviously doubted it very much. "To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it. But don't you think that life with a statistician would be rather, shall we say, humdrum?"

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she had discovered below Edward's numerical veneer.

"It's not the figures themselves," she said finally. "it's what you do with them that matters."

(K.A.C. Manderville, The Undoing of Lamia Gurdleneck)

Example: z-test for the normal location problem

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, where σ_0^2 is known. Let μ_0, μ_1 be given and $\mu_1 > \mu_0$. The optimal size α test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ rejects H_0 when

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} > z_\alpha.$$

- *[Show this using the Neyman-Pearson Lemma.]*

Numerical example

- $\mu_0 = 5$, $\mu_1 = 6$, $\sigma_0 = 1$, $\alpha = 0.05$, $n = 9$.
- Observed data: $X = (5.1, 5.5, 4.9, 5.3, 5.2, 5.3, 5.7, 5.0, 4.8)$, so $\bar{X} = 5.2$.
- $Z = 0.6 < z_{0.05} \approx 1.645$.
- Not enough evidence against H_0 (at significance level 0.05).

P-values and significance tests (skip and see Lecture 9)

When $\Theta_0 = \{\theta_0\}$, the result of a LR test can be summarized by the P-value:

$$P = \mathrm{P}_{\theta_0}(\Lambda(\tilde{X}) \geq \Lambda(X) \mid X),$$

where $\tilde{X}$ is an IID copy of X (so $\tilde{X} \sim \mathrm{p}(x; \theta_0)$ and $\tilde{X}$ is independent of X).

- This is a statistic—the probability of observing a more extreme test statistic under the null hypothesis.
- The test rejects H_0 at significance level α iff $P \leq \alpha$. *[Why?]*
- *[What's the p-value for the z-test?]*

Proposition (probability integral transform)

Suppose $\Lambda(X)$ has a continuous distribution under $\theta = \theta_0$. Then $P \sim \mathrm{Unif}[0, 1]$.

- *[Proof.]*
- *[What happens if the distribution of Λ has jumps?]*

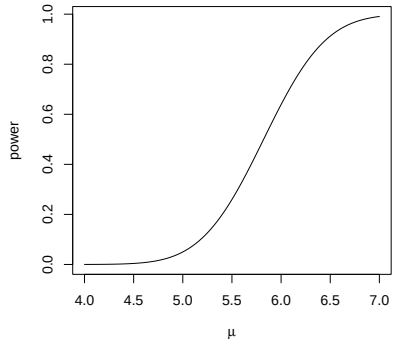
Outline

Power function for the normal location problem

Recall that the power function of a test $T(X) \in \{0, 1\}$ is $\beta(T; \theta) = P_\theta(T(X) = 1)$.

- The z-test rejects $H_0 : \mu = \mu_0$ if $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 > z_\alpha$.
- The power function is $\beta(\mu) = 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}\right)$. *[Derive this.]*

```
n <- 4; sigma0 = 1; mu0 <- 5; alpha <- 0.05
power <- 1 - pnorm(qnorm(1 - alpha) + sqrt(n) * (mu0 - x) / sigma0)
```



Optimality of hypothesis tests

Consider any disjoint $\Theta_0, \Theta_1 \subset \Theta$.

Definition

$T : \mathcal{X} \to \{0, 1\}$ is the uniformly most powerful (UMP) size α test for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, if

1. $\sup_{\theta \in \Theta_0} \beta(T; \theta) = \alpha$;
2. for any other test $\tilde{T}$ with size $\leq \alpha$, we have $\beta(T; \theta) \geq \beta(\tilde{T}; \theta)$ for all $\theta \in \Theta_1$.

Examples

1. The z-test is UMP for testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ in the normal location problem. *[Why?]*
2. *[*Monotone likelihood ratio and one-parameter exponential family.]*

Generalized likelihood ratio test

Let Θ_0, Θ_1 be disjoint subsets of Θ.

Definition

The (generalized) likelihood ratio statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_0 : \theta \in \Theta_1$ is

$$\Lambda(X) = \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} \mathrm{p}(X; \theta)}{\sup_{\theta \in \Theta_0} \mathrm{p}(X; \theta)}.$$

The size α (generalized) likelihood ratio test rejects H_0 if $\Lambda > c_\alpha$, where c_α satisfies

$$\sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\Lambda(X) > c_\alpha) = \alpha.$$

- N-P lemma: optimality when $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.
- To find $\Lambda(X)$, need to compute two (restricted) MLEs.
- c_α is easier to find if there is a "worst-case null distribution" for all α.

*Interlude: Pareto efficiency

- Many real-life problems involve trading off different objectives.
- A decision rule is Pareto efficient or admissible if there is no other rule that improves all objectives.

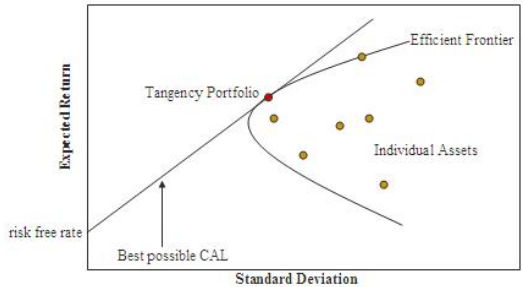
Example: portfolio optimization

Let $X_1, \ldots, X_n$ be the return of n stocks. The optimal portfolio solves

$$\begin{aligned} \text{minimize} \quad & w^T \operatorname{Cov}(X) w \\ \text{subject to} \quad & w^T \operatorname{E}(X) = \mu, \\ & \sum_{i=1}^{n} w_i = 1. \end{aligned}$$

- This is a quadratic program and can be easily solved using Langrange multiplier.

*Interlude: efficient frontier for portfolio

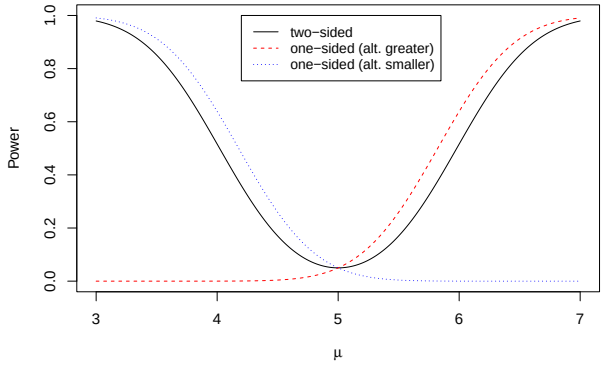


- CAL = Capital Allocation Line.
- Seminal work by Harry Markowitz (1952), which won him a Nobel Prize in 1990. Later discovered that the idea can be found in work by Bruno de Finetti in 1940.
- Part II Stochastic Financial Models.

Two-sided normal location problem: two-sided z-test

$X_1, \ldots, X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, σ_0^2 is known. Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ for given μ_0.

- The generalized LRT rejects H_0 if $|Z| > z_{\alpha/2}$ where $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0$. *[Why?]*
- *This is the UMP unbiased test. The UMP test does not exist. *[Why?]*



Asymptotic distribution of generalized likelihood ratio tests

- Loosely speaking, the degrees of freedom of a statistical model is its number of "free parameters". *[Examples.]*

Wilks' theorem

Suppose $\Theta \subseteq \mathbb{R}^d$ is open and Θ_0 is a d_0-dimensional linear subspace of Θ. Under regularity conditions, we have

$$2 \log \Lambda(X) \xrightarrow{d} \chi^2_{d-d_0} \text{ as } n \to \infty \text{ under } H_0.$$

- This gives a universal rejection threshold: $c_\alpha \approx \chi^2_{d-d_0}(\alpha)$ for large n.
- The χ^2-distribution is exact for the two-sided normal location problem.
- Proof in Part II Principles of Statistics.

Outline

P-values and significance tests

- Often, a test (e.g. LRT) rejects H_0 if $\Lambda(X) > c_\alpha$ for some $\Lambda(X) \in \mathbb{R}$, and c_α is chosen so that $\sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\Lambda(X) > c_\alpha) = \alpha$.
- We can then summarize the test using a statistic called the "p-value":

$$P = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\Lambda(\tilde{X}) \geq \Lambda(X) \mid X),$$

  where $\tilde{X}$ is an IID copy of X (so $\tilde{X} \sim \mathrm{p}(x; \theta_0)$ and $\tilde{X}$ is independent of X).
- This is the probability of observing a more extreme test statistic under H_0.
- The test rejects H_0 at level α iff $P \leq \alpha$. *[Why?]*
- *[What's the p-value for the one-sided and two-sided z-tests?]*

Proposition (probability integral transform)

Suppose $\Theta_0 = \{\theta_0\}$ and $\Lambda(X)$ is continuous under $\theta = \theta_0$. Then $P \sim \mathrm{Unif}[0, 1]$.

- *[Proof.]*
- *[What happens if the distribution of Λ has jumps?]*

*P-value wars

- P-value is the single most used number in scientific studies to quantify strength of evidence.
- But p-values are often misused: hunting for significance ("p-hacking").
- There have been "p-value wars" in the last decade:
  - The ASA Statement on p-Values: Context, Process, and Purpose (Wasserstein and Lazar, 2016).
  - The ASA presidents task force statement on statistical significance and replicability (Benjamini et al., 2021).
- "It's not the figures themselves, it's what you do with them that matters".

Duality between simple tests and confidence intervals

Suppose $X \sim \mathrm{p}(x;\theta)$, $\theta \in \Theta$, θ is unknown.

Theorem

Consider $T : \Theta \times \mathcal{X} \to \{0,1\}$ and $I : \mathcal{X} \to 2^{\Theta}$ that satisfies

$$I(X) = \{\theta : T(\theta, X) = 0\} \quad \text{or equivalently} \quad T(\theta, X) = 1_{\{\theta \notin I(X)\}}.$$

Then the following statements are equivalent:

1. $T(\theta_0, X)$ is a size α test of $H_0 : \theta = \theta_0$ for all $\theta_0 \in \Theta$;
2. $I(X)$ is a $(1-\alpha)$-confidence set for θ.

- *[Proof.]*
- *[Alternative version with p-value.]*
- *[Example: two-sided normal location problem.]*

Motivating example for goodness-of-fit: Gregor Mendel's pea experiments

- From 1856 to 1863, Mendel produced around 29,000 garden pea plants from controlled crosses and registered their phenotypes.
- In one experiment, he crossed $n = 566$ round yellow peas with green wrinkled peas.

	Round Yellow	Round Green	Wrinkled Yellow	Wrinkled Green
Count	315	108	101	32
Proportion	0.557	0.191	0.178	0.057
Theory	0.563	0.188	0.188	0.063

- According to what we now call Mendel's laws of inheritance, the ratio is 9:3:3:1 in theory. Is this a good fit to the data?

Goodness-of-fit testing

Suppose $(N_1, \ldots, N_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$.

- Interested in testing $H_0 : p_i = p_{0i}, i = 1, \ldots, k$ for a given vector p_0.
- The generalized log-likelihood ratio statistic is given by

$$2 \log \Lambda = 2 \sum_{i=1}^{k} N_i \log \left(\frac{N_i}{n p_{0i}} \right) = 2 \sum_{i=1}^{k} O_i \cdot \log \left(\frac{O_i}{E_i} \right),$$

  where $O_i = N_i$ is the "observed count" and $E_i = n p_{0i}$ is the "expected count".
- When $O_i / E_i \approx 1$, this can be approximated by Pearson's statistic

$$2 \log \Lambda \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

- *[Derive these and the associated χ^2-test.]*
- *[Rule of thumb. *Demonstration in R using Mendel's data.]*

*Interlude: The Mendel-Fisher controversy

- Fisher (1936) concluded that "the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations."

Experiments	Expectation	χ^2	Probability of exceeding deviations observed
3 : 1 ratios	7	2.1389	0.95
2 : 1 ratios	8	5.1733	0.74
Bifactorial	8	2.8110	0.94
Gametic ratios	15	3.6730	0.9987
Trifactorial	26	15.3224	0.95
Total	64	29.1186	0.99987
Illustrations of plant variation	20	12.4870	0.90
Total	84	41.6056	0.99993

- Table 5 from Pires and Branco (2010), who also offered an explanation using statistical models in which Mendel only stopped the experiments when the data are "close enough" to his theory.

Outline

Composite null

E B Ford (1971) recorded wing patterns of scarlet tiger moth:

Phenotype	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	Total
Number	1469	138	5	1612

- Hardy-Weinberg equilibrium: $\mathrm{P}(AA) = \theta^2$, $\mathrm{P}(Aa) = 2\theta(1-\theta)$, $\mathrm{P}(aa) = (1-\theta)^2$ under random mating and no evolutionary influences.
- *[Develop the generalized likelihood ratio test for H-W equilibrium.]*
- *[*Demonstration in `R` using Ford's data.]*

Testing independence in 2-way contingency tables

Suppose $(X_1, Y_1), \ldots (X_n, Y_n) \overset{IID}{\sim} \mathrm{P}$, $X_1 \in \{1, \ldots, r\}$, $Y_1 \in \{1, \ldots, c\}$. Wish to test H_0: X_1 and Y_1 are independent: $\mathrm{P}(X_1 = x, Y_1 = y) = \mathrm{P}(X_1 = x)\,\mathrm{P}(Y_1 = y)$.

- Data can be summarized by a contingency table $N_{xy} = \sum_{i=1}^{n} 1_{\{X_i = x, Y_i = y\}}$.

Example

500 people with recent car changes were asked about their previous and new cars.

		New car		
		Large	Medium	Small
Old car	Large	56	52	42
	Medium	50	83	67
	Small	18	51	81

χ^2-test for independence

- We can model the contingency table by

$$(N_{11}, \ldots, N_{1c}, \ldots, N_{r1}, \ldots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, \ldots, p_{1c}, \ldots, p_{r1}, \ldots, p_{rc}).$$

- $H_0 : p_{xy} = p_{x+}p_{+y}$ (with $\sum_x p_{x+} = 1 = \sum_y p_{+y}$, $p_{x+}, p_{y+} \geq 0$).
- H_1: p_{xy} is unrestricted other than $\sum_{x,y} p_{xy} = 1, p_{xy} \geq 0$.
- *[Derive the χ^2-test for independence. What's its degrees of freedom?]*
- *[*Demonstration in `R` (including Pearson's residuals).]*

Tests of homogeneity

- Consider the car change dataset again. Will my test be different if I know the row totals are fixed in sampling?

		New car			
		Large	Medium	Small	Total
Old	Large	56	52	42	150
car	Medium	50	83	67	200
	Small	18	51	81	150
	Total	124	186	190	500

- Model: $(N_{x1}, \ldots, N_{xc}) \overset{ind.}{\sim} \text{Multinomial}(n_x; p_{1|x}, \ldots, p_{c|x})$, $x = 1, \ldots, r$.
- *[How is this different and related to the IID model?]*
- Interested in testing H_0: $p_{y|x}$ does not depend on x.
- The corresponding generalized LRT is exactly the same as that for independence testing. *[ES2]*

*Permutation tests

- A drawback of the χ^2-test is that the χ^2 null distribution is only approximate.
- For independence testing in 2-way tables, we can use the permutation null instead. Suppose $(X_1, Y_1), \ldots (X_n, Y_n)$ are IID. If X_1 and Y_1 are independent, then

$$(X_1, Y_1, \ldots, X_n, Y_n) \stackrel{d}{=} (X_1, Y_{\pi(1)}, \ldots, X_n, Y_{\pi(n)})$$

  for any permutation π of $\{1, \ldots, n\}$.
- Alternatively, this amounts to generating random tables with given marginal totals.
- *[What is the p-value of the permutation test?]*
- When applying `chisq.test` in `R`, this can be obtained by setting `simulate.p.value=TRUE`.

*Interlude: Lady tasting tea

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup... Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist... Her task is to divide the 8 cups into two sets of 4.
(R A Fisher. The Design of Experiments. 1935)

- Fisher uses this experiment to articulate why randomization provides a "reasoned basis" for causal inference. This marks a key moment in the 20th-century scientific revolution from determinism (clockwork universe/Laplace's demon) to a statistical view.
- Further reading:
  - I. Hacking. (1988). Telepathy: Origins of Randomization in Experimental Design.
  - My blog post: The origin of randomization.

*Fisher's exact test

The outcome of Fisher's experiment can be summarized in a 2×2 table:

$$N_{ij} = \text{Number of cups made by } i \text{ with guess } j, \quad i, j \in \{\text{milk first}, \text{tea first}\}$$

- By design, $N_{++} = 8$ and $N_{1+} = N_{2+} = N_{+1} = N_{+2} = 4$. So only one "degree of freedom". *[Why?]*
- H_0: observe a random table (the lady guessed randomly). Reject if N_{11} is large.
- The p-value $P(N_{11})$ is the probability of observing the same or more extreme N_{11}.

$$P(4) = 1/\binom{8}{4} = 1/70 = 1.4\%, \quad P(3) = (1+\binom{4}{3}\binom{4}{1})/\binom{8}{4} = 17/70 = 24.3\%.$$

  This is called the hypergeometric distribution.
- *[Why is this a permutation test?]*
- *[*Demonstration of `fisher.test` in R.]*

Outline

Multivariate normal

Let $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ be a random vector and denote $\mu = \mathrm{E}(X)$ and $\Sigma = \mathrm{Cov}(X)$. Let $A \times \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be fixed.

- We know $\mathrm{E}(AX + b) = A\mu + b$ and $\mathrm{Cov}(AX + b) = A\Sigma A^T$. *[Why?]*

Definition

We say X has a multivariate normal distribution and if $a^T X$ has a (univariate) normal distribution for all fixed $a \in \mathbb{R}^n$.

- The MGF is $M_X(a) = \mathrm{E}(e^{a^T X}) = e^{a^T \mu + a^T \Sigma a/2}$. *[Why?]*
- By uniqueness of MGF, the distribution of X is determined by its mean and covariance matrix, so we write $X \sim \mathrm{N}(\mu, \Sigma)$.
- The density function of X is given by

$$\mathrm{p}(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} e^{(x-\mu)^T \Sigma^{-1} (x-\mu)/2}. \quad \textit{[Why?]}$$

- If $X \sim \mathrm{N}(\mu, \Sigma)$, then $AX + b \sim \mathrm{N}(A\mu + b, A\Sigma A^T)$. *[ES3]*

Normal location problem with unknown variance

Suppose $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

- Recall $\bar{X} = \sum_i X_i/n$ and $S_{XX} = \sum_i (X_i - \bar{X})^2$.

Theorem

$\bar{X}$ and S_{XX} are independent, $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, and $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$. *[Proof.]*

Student's t-test

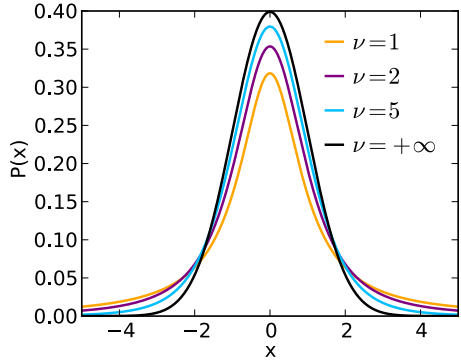
- Suppose Z and Y are independent, $Z \sim N(0,1)$ and $Y \sim \chi^2_k$. We say $T = \frac{Z}{\sqrt{Y/k}}$ follows the t-distribution with k degrees of freedom and write $T \sim t_k$.
- Student's t-statistic is given by

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{S_{XX}/(n-1)}} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}, \text{ where } \hat{\sigma}^2 = S_{XX}/(n-1).$$

- *[Show that $T \sim t_{n-1}$ when $\mu = \mu_0$. "Standard error".]*
- *[State the one-sided and two-sided t-tests, *which are the UMP unbiased tests.]*

*Some facts about the t-distribution

- PDF of t_k is given by $\mathrm{p}(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\Gamma(\frac{k}{2})}(1 + x^2/k)^{-(k+1)/2}$.



- t_1 = Cauchy distribution; $t_k \xrightarrow{d} \mathrm{N}(0,1)$ when $k \to \infty$. ($\nu = k$ in figure above.)
- Tail behaves like $x^{-(k+1)}$, an instance of power law. $\mathrm{E}(|X|^m) < \infty$ for $0 < m < k$.

*Interlude: History of the *t*-test

- At the turn of the 20th century, brewers at the Guinness factory in Dublin were interested in testing the amount of soft resins in different batches of hop flowers, which impart a bitter flavour and act as a natural preservative.
- W S Gosset was the head experimental brewer at Guinness and developed the *t*-test. In 1908, he published his result under the pen name *Student*. (For confidentiality reasons, Guinness allowed its scientists to publish research on condition that they do not mention beer, Guinness, or their own surname.)

Orthogonal projections (skip; see Lecture 13)

- $X \sim \mathrm{N}(0, I_n)$ is isotropic: $UX \sim \mathrm{N}(0, I_n)$ for all orthogonal matrix U.

Definition/Proposition

Consider $P \in \mathbb{R}^{n \times n}$. The following are equivalent:

1. P is an (orthogonal) projection matrix.
2. P is symmetric ($P^T = P$) and idempotent ($P^2 = P$).
3. $P = UU^T$, where columns of $U \in \mathbb{R}^{n \times \mathrm{rank}(P)}$ form an orthonormal basis for the column space of P.

[Proof. Why is $(I - P)$ also a projection? Why is it true that $\mathrm{rank}(P) = \mathrm{tr}(P)$?]

Theorem

Suppose $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ and P is a projection matrix. Then

1. $P\epsilon \sim \mathrm{N}(0, \sigma^2 P)$ and $(I - P)\epsilon \sim \mathrm{N}(0, \sigma^2(I - P))$ are independent.
2. $\|P\epsilon\|^2/\sigma^2 \sim \chi^2_{\mathrm{rank}(P)}$.

[Proof. Specialization to t-test for normal location problem.]

Outline

Motivating example

Groups of 25 males kept with	Mean life (days)	$\hat{\sigma} = \sqrt{S_{XX}/(n-1)}$
no companions	63.56	16.4522
1 uninterested female	64.80	15.6525
1 interested female	56.76	14.9284
8 uninterested females	63.36	14.5398
8 interested females	38.72	12.1021

Partridge and Farquhar (1981) "Sexual activity reduces lifespan of male fruitflies".

- *[Construct 95% confidence intervals for each row. *Demonstration in R.]*
- Motivation for ANOVA: testing the equality of means of two or more rows.

One-way analysis of variance (ANOVA)

Given $Y_{ij} = \mu_i + \epsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, k$, we wish to test $H_0 : \mu_1 = \cdots = \mu_k$.

- $k \geq 2$ number of groups. $n_i \geq 1$ observations in group i. Let $\sum_{i=1}^{k} n_i = n$.
- We assume $\epsilon_{ij} \overset{IID}{\sim} \mathrm{N}(0, \sigma^2)$. *[What is the distribution of Y?]*

F-test

- The generalized LR statistic is an increasing function of SSA/SSE, where

$$\mathrm{SSA} = \sum_{i=1}^{k} n_i (\bar{Y}_{i+} - \bar{Y}_{++})^2, \quad \mathrm{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2. \text{ [Why?]}$$

- *[Show that SSA and SSE and independent.]*
- If $S_1 \sim \chi^2_{d_1}$ and $S_2 \sim \chi^2_{d_2}$ are independent, then we say $R = (S_1/d_1)/(S_2/d_2)$ follows the F-distribution with degrees of freedom d_1 and d_2 and write $R \sim F_{d_1, d_2}$.
- The F-test rejects H_0 if $(\mathrm{SSA}/(k-1))/(\mathrm{SSE}/(n-k)) > F_{k-1, n-k}(\alpha)$.
- *[Why is this a size-α test?]*

ANOVA table

Source of variation	Sum of squares	DF	Mean squares	F stat.
Between groups	SSA	$k-1$	$\text{MSA} = \text{SSA}/(k-1)$	MSA/MSE
Within groups	SSE	$n-k$	$\text{MSE} = \text{SSE}/(n-k)$	
Total	SST	$n-1$		

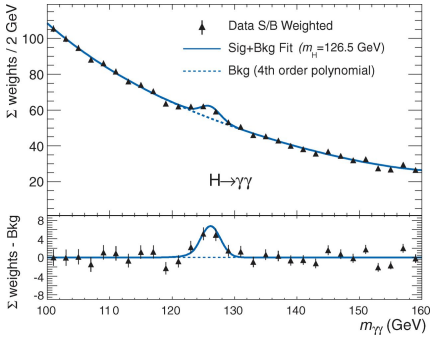
$$\text{SST} = \text{SSA} + \text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{++})^2.$$

Numerical example

- H_0: equal mean in three control groups in Partridge and Farquhar (rows 1, 2, 4).
- $\bar{X} = 63.91$. SSE $= 17449.92$. SSA $= 30.427$.
- F statistic is 0.0628, p-value is 0.939.
- *[*Calculations in R.]*

*Interlude: Discovery of Higgs boson

- The Higgs Discovery Explained | CERN.



- A simplified model: $N \sim \text{Poisson}(\beta + \kappa\mu)$. $H_0 : \mu = 0$ vs. $H_1 : \mu = 1$.
  - β and κ are expected background and Higgs boson counts, respectively.
- The physicists essentially used a (generalized) LRT.
- Many statistical issues such as model misspecification and estimation of HB mass (van Dyk, 2014: The Role of Statistics in the Discovery of a Higgs Boson).

Normal linear model

We have data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$.

Definition

The normal linear model assumes

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n \overset{IID}{\sim} \mathrm{N}(0, \sigma^2)$ and are independent of $(X_1, \ldots, X_n)$.

- *[Matrix form. What is the distribution of $Y = (Y_1, \ldots, Y_n)$?]*

Example 1: ANOVA as a normal linear model

- *[One-way ANOVA as nested linear models. What are the design matrices?]*
- *[SSA and SSE as orthogonal projections.]*

Example 2: Simple linear regression

- $Y_i = \alpha + \beta X_i + \epsilon_i, \;\; i = 1, \ldots, n$. *[What is the design matrix?]*

Outline

Linear models

"Design matrix" $X \in \mathbb{R}^{n \times p}$ and "response" $Y \in \mathbb{R}^n$. Each row is a single observation.

- Some columns of X can be fixed. For example, a column of 1 models "intercept".
- "Linear model" can mean many different things. We will only discuss two variants.

Definition

Consider unknown parameters $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$.

1. The (homoscedastic) normal linear model:

$$Y \mid X \sim \mathrm{N}(X\beta, \sigma^2 I_n).$$

2. The (homoscedastic) linear conditional expectation model:

$$\mathrm{E}(Y \mid X) = X\beta, \;\; \mathrm{Var}(Y \mid X) = \sigma^2 I_n.$$

[Express these models using $\epsilon = Y - X\beta$ and for each observation (X_i, Y_i).]

Least squares estimator

- For the normal linear model, the log-likelihood function is given by

$$l(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2 + \text{const.} \quad \textit{[Derive this.]}$$

- The MLE of β is given by the (ordinary) least squares estimator

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - X_i^T\beta)^2.$$

- Will always assume X has full column rank. Then $\hat{\beta} = (X^TX)^{-1}X^TY$. *[Why?]*
- *[What is the least squares estimator for simple linear regression?]*

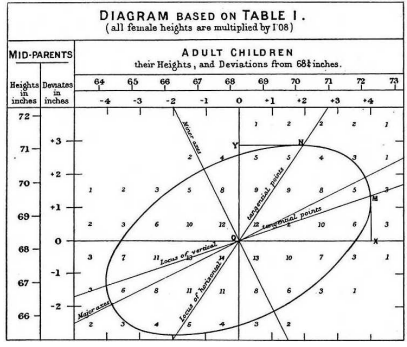
Gauss-Markov Theorem ("Best Linear Unbiased Estimator/BLUE")

Let $\tilde{\beta}$ be another unbiased estimator of β in the linear conditional expectation model that is linear in Y. Then $\text{Var}(a^T\hat{\beta}) \leq \text{Var}(a^T\tilde{\beta})$ for all $a \in \mathbb{R}^p$.

- *[Proof. Why is $\hat{\beta}$ unbiased?]*

*Interlude: Regression towards mean/mediocrity

- Least squares/linear models date back to Newton/Legendre/Gauss/Quetelet, but now often known as "linear regression" due to influential work by Francis Galton.



- "vertical tangential": OLS children on parents; "major axes": total least squares.
- Galton coined the term "eugenics". It originated as a progressive social movement in 19th century, but now basically means scientific racism.

Orthogonal projections

The least squares problem can be equivalently written as

$$\hat{\mu} = \arg \min_{\mu \in \text{colspan}(X)} \|Y - \mu\|^2.$$

- The solution is given by the fitted values: $\hat{\mu} = X\hat{\beta} = PY$ for $P = X(X^TX)^{-1}X^T$.
- The vector of residuals is given by $R = Y - \hat{\mu} = (I - P)Y$.

Definition

We say $P \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix if it is symmetric ($P^T = P$) and idempotent ($P^2 = P$).

Proposition

$P \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix if and only if $P = UU^T$, where columns of $U \in \mathbb{R}^{n \times \text{rank}(P)}$ form an orthonormal basis for the column space of P.
[Proof.]

More on orthogonal projections

Proposition

Suppose $P \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix. Then

1. The eigenvalues of P is either 0 or 1.
2. $\text{rank}(P) = \text{tr}(P)$.
3. $I - P$ is also an orthogonal projection matrix.
4. $\|Y\|^2 = \|PY\|^2 + \|(I - P)Y\|^2$ for any $Y \in \mathbb{R}^n$.

Application to linear model

- The projection matrix $P = X(X^T X)^{-1} X^T$ is invariant to scaling and translating the columns of X (the latter requires an intercept term in the model).
- *[Sample mean and variance as projections.]*

Outline

Review: Normal linear model

- Recall that the normal linear model assumes $Y \mid X \sim \mathrm{N}(X\beta, \sigma^2 I_n)$.
  - We will usually treat X as fixed and drop the conditioning.
- The log-likelihood function is given by

$$l(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2 + \text{const.}$$

- The MLE is given by (assuming X has full column rank):

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\|Y - X\hat{\beta}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} R_i^2 = \text{"Residual Sum of Squares"}/n.$$

- $\hat{\beta}$ is BLUE even if noise is not normal.

*Interlude: Latin square

- A Latin square is an $n \times n$ array with n letters so that no letter appears more than once in any row or column. (Variations: Sudoku, eight queens puzzle.)

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

A	A	B	B
A	A	B	B
A	A	B	B
C	C	D	C

A	A	B	B
A	A	B	B
C	C	D	D
D	C	D	D

- No. of LS: 1 ($n = 2, 3$), 4 ($n = 4$), 56 ($n = 5$), 9408 ($n = 6$), 16942080 ($n = 7$).

Experimental design

- Think about a plot of apple trees and four different fertilizers (A, B, C, D).
- Yield in (i, j) is $Y_{ij} = \mu_i + \lambda_j + \theta_{F_{ij}} + \epsilon_{ij}$, $F_{ij} \in \{A, B, C, D\}$, $\epsilon_{ij} \overset{IID}{\sim} \mathrm{N}(0, \sigma^2)$.
- Let $\hat{\theta}_F$ be the average yield of all plots with F. Then $\hat{\theta}_A - \hat{\theta}_B$ is the BLUE for $\theta_A - \theta_B$, and Latin squares minimize the max variance of $\hat{\theta}_A - \hat{\theta}_B$, $\hat{\theta}_B - \hat{\theta}_C$, etc.

Distribution of the MLE

Lemma

Suppose $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ and P is an orthogonal projection matrix. Then

1. $P\epsilon \sim \mathrm{N}(0, \sigma^2 P)$ and $(I - P)\epsilon \sim \mathrm{N}(0, \sigma^2(I - P))$ are independent.
2. $\|P\epsilon\|^2/\sigma^2 \sim \chi^2_{\mathrm{rank}(P)}$.

Theorem

Under the normal linear model (with fixed X of full column rank),

1. $\hat{\beta}$ and $\hat{\sigma}^2_{\mathrm{MLE}}$ are independent.
2. $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(X^T X)^{-1})$.
3. $n\hat{\sigma}^2_{\mathrm{MLE}}/\sigma^2 \sim \chi^2_{n-p}$.

[Prove the above results.]

[Application to Student's t-test for the normal location problem.]

Some implications

1. $\hat{\sigma}^2_{\text{MLE}}$ is a biased estimator. It is more common to use the unbiased

$$\hat{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2.$$

2. The (two-sided) t-test rejects $H_0 : \beta_j = \beta_{0j}$ (vs. $H_1 : \beta_j \neq \beta_{0j}$) if

$$\frac{|\hat{\beta}_j - \beta_{0j}|}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}} > t_{n-p}(\alpha/2).$$

[Why is this a level-α test? What is the corresponding confidence interval]

3. The F-test rejects $H_0 : \beta = \beta_0$ (vs. $H_1 : \beta \neq \beta_0$) if

$$\frac{\|X(\hat{\beta} - \beta_0)\|^2/p}{\hat{\sigma}^2} > F_{p,n-p}(\alpha).$$

[Why is this level-α? What is the shape of the corresponding confidence set?]

Example: Two-sample t-test

Suppose $A_1, \ldots, A_n \overset{IID}{\sim} \mathrm{N}(\mu_1, \sigma^2)$ and $B_1, \ldots, B_m \overset{IID}{\sim} \mathrm{N}(\mu_2, \sigma^2)$ are independent. The parameters $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

- *[Express this as a normal linear model.]*
- *[What is the two-sided t-test for $H_0 : \mu_1 = \mu_2$?]*
- *[What is the ANOVA F-test for $H_0 : \mu_1 = \mu_2$?]*
- *[Show that these two tests are equivalent.]*

Outline

Nested projections

- Recall $P \in \mathbb{R}^{n \times n}$ is an orthogonal projection if it is symmetric and idempotent.
- In least squares problems, $P = X(X^T X)^{-1} X^T$ projects onto colspan(X) and can be written as $P = UU^T$ where $U \in \mathbb{R}^{n \times p}$ is an orthonormal basis of X.

Consider the partition $X = (X_0 \ X_1)$, $X_0 \in \mathbb{R}^{n \times p_0}$ and $X_1 \in \mathbb{R}^{n \times (p - p_0)}$ $(0 < p_0 < p)$.

- Define $P = X(X^T X)^{-1} X$ and $P_0 = X_0 (X_0^T X_0)^{-1} X_0$.
- Key geometric result: $PP_0 = P_0 P = P_0$. *[Proof.]*

Theorem

Assume $X \in \mathbb{R}^{n \times p}$ has full column rank. For $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, we have

1. $P_0 \epsilon$, $(P - P_0)\epsilon$, $(I - P)\epsilon$ are independent.
2. $\|P_0 \epsilon\|^2 \sim \chi^2_{p_0}$, $\|(P - P_0)\epsilon\|^2 \sim \chi^2_{p - p_0}$, $\|(I - P)\epsilon\|^2 \sim \chi^2_{n - p}$.

[Proof.]

[Remark: These results only require colspan(X_0) to be a linear subspace of colspan(X).]

Testing nested models

For simplicity of exposition, we will treat X as fixed. Let $\mu = \mathrm{E}(Y)$.

Model	Hypothesis	Design	Fitted values
Saturated	$\mu \in \mathbb{R}^n$	I_n	$\hat{\mu} = Y$
Full	$\mu \in \mathrm{colspan}(X)$	X	$\hat{\mu} = PY$
Sub/Null	$\mu \in \mathrm{colspan}(X_0)$	X_0	$\hat{\mu} = P_0 Y$

Theorem (general linear hypothesis tests)

Consider the testing problem H_0 : sub-model vs. H_1 : full model (minus sub-model). Then the size α generalized LRT rejects H_0 if

$$\frac{\|(P - P_0)Y\|^2/(p - p_0)}{\|(I - P)Y\|^2/(n - p)} > F_{p-p_0,n-p}(\alpha).$$

- *[Prove this. How can H_0 be expressed as a constraint on β?]*

Analysis of variance

The F-test for nested models generalizes many tests before:

1. Student's two-sided t-test for normal location problem.
2. The one-way analysis of variance.
3. The two-sided t-test for $H_0 : \beta_j = \beta_{0j}$ in normal linear models.

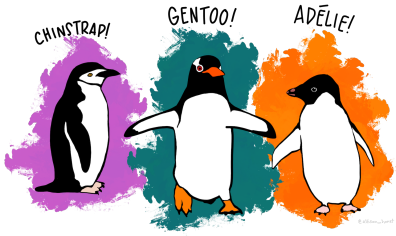
*Sequential ANOVA

Let $M_0 \subset M_1 \subset \cdots \subset M_k = \mathbb{R}^n$ be a nested sequence of linear spaces. Let p_j be the dimension of M_j and P_j be the projection matrix onto colspan(M_j).

- Let $\mu = \mathrm{E}(Y)$. The F-test for $H_{j-1} : \mu \in M_{j-1}$ vs. $H_j : \mu \in M_j$ rejects H_{j-1} if

$$\frac{\|(P_j - P_{j-1})Y\|^2/(p_j - p_{j-1})}{\|(P_{j+1} - P_j)Y\|^2/(p_{j+1} - p_j)} > F_{p_j - p_{j-1}, p_{j+1} - p_j}(\alpha).$$

*Example in R



- Load the `penguins` dataset in `palmerpenguins` package.
- `?penguins`.
- `summary`, `subset`, `boxplot`, `table`.
- `lm`, `summary.lm`, `anova`.
- `plot`, `abline`.
- Yule-Simpson paradox.

Outline

Two-sample testing problems

Suppose $A_1, \ldots, A_n \overset{IID}{\sim} F$ and $B_1, \ldots, B_m \overset{IID}{\sim} G$. Let $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_m)$.
Are F and G similar?

Different formulations

1. A and B are independent, $F = \mathrm{N}(\mu_1, \sigma^2)$, $G = \mathrm{N}(\mu_2, \sigma^2)$, $H_0 : \mu_1 = \mu_2$.
   - We can use the two-sample t-test or equivalently the ANOVA F-test.
2. *A and B are independent, $F = \mathrm{N}(\mu_1, \sigma_1^2)$, $G = \mathrm{N}(\mu_2, \sigma_2^2)$, $H_0 : \mu_1 = \mu_2$
   - *Behrens-Fisher problem. A popular, approximate solution is Welch's t-test.
3. *A and B are independent, no assumption on the form of F or G, $H_0 : F = G$.
   - Can use permutation t-test, which is asymptotically equivalent to the t-test in 1.

*[*Demonstration in R.]*

Paired t-test

Sometimes the observations were paired.

Example

Should supervisors ask students to turn in their work 1 or 2 days before the supervision? A good experimental design randomizes the requirement within each pair of supervisees.

Model and solution

Let (A_i, B_i) be the exam results in pair i.

$$A_i \sim \mathrm{N}(\mu_1 + \alpha_i, \sigma^2), \;\; B_i \sim \mathrm{N}(\mu_2 + \alpha_i, \sigma^2), \;\; i = 1, \ldots, n,$$

and are all independent.

- $D_i = A_i - B_i \overset{IID}{\sim} \mathrm{N}(\mu_1 - \mu_2, 2\sigma^2)$.
- Thus can apply Student's t-test for $D_1, \ldots, D_n$.
- *[How can this be set up as a normal linear model? *Demonstration in R.]*

*Sample size planning

When designing experiments, it is often useful to estimate the sample size required to achieve certain power (under a given alternative hypothesis).

Example

A clinical trial randomly assigns n patients to placebo and n to a new therapy. We are interested in testing whether the new therapy lowers blood pressure.

- What is the minimum n required so that the null hypothesis is rejected at level α with probability β when the treatment effect is δ?

Power calculation

- Suppose $Z \sim \mathrm{N}(0,1)$ and $Y \sim \chi^2_k$ are independent. We say $\frac{Z+\mu}{\sqrt{Y/k}}$ follows the noncentral t-distribution with k d.o.f. and noncentrality parameter $\mu \in \mathbb{R}$.
- *[Derive the power function of two-sample t-test.]*

"Stats Lab courses" in Part II

Michaelmas

- Statistical Modelling: various extensions to the normal linear model; statistical computing with R.
- Principles of Statistics: likelihood principle; basic asymptotic statistics; Bayesian inference and decision theory; basic nonparametric statistics and MCMC.
- Stochastic Financial Models: utility and mean-variance analysis; dynamic programming; introduction to martingales and Brownian motions; Black-Scholes model for option pricing.
- Probability and Measure: rigorous treatment of the foundation.

Lent

- Mathematics of Machine Learning: statistical learning theory; empirical risk minimization; popular machine learning methods.
- Applied Probability: continuous-time Markov chains; Poisson processes and renewal processes; applications to queueing theory.