Data analysis

Ben Graham

MA930, University of Warwick

October 26, 2015

p-value examples

- p-value: statistic P such that $\mathbb{P}(P \leq t) \leq t$ for $t \in [0, 1]$.
- Random sample $X_1, \ldots, X_n$ iidrv Exponential(θ), mean $\mu := 1/\theta$ unknown
- Under $H_0 : \mu = 1$, sample mean has the Gamma distribution with mean 1, shape n
  - Let $Q = F_{H_0}(\bar{X}) \in [0, 1]$ — uniform under H_0
- $H_1 : \mu < 1$
  - Reject smallvalues of Q
  - $P = Q$
- $H_1 : \mu \neq 1$
  - Reject large values of Q
  - $P = 1 - Q$
- $H_1 : \mu \neq 1$
  - Reject small and large values of Q
  - $P = 1 - |2Q - 1|$

Hypothesis testing

- Example: "Lady tasting tea"
- Experiments by Ronald Fisher and Muriel Bristol
- 8 cups of tea: 4 tea into cup first, 4 milk first

#Successes	Count
0	1
1	$16 = \binom{4}{1}\binom{4}{3}$
2	$36 = \binom{4}{2}\binom{4}{2}$
3	$16 = \binom{4}{3}\binom{4}{1}$
4	1

- $\sum count = 70 = \binom{8}{4}$
- Null hypothesis: no ability to taste difference — construct p-value
- Under H_0, $\mathbb{P}(4 \text{ successes} \mid H_0) = 1/70$
- Muriel got 4/4: Reject H_0

Power calculations

- Before doing an experiment, check it can do what you want
- Example: Testing a coin for bias: $X \sim$Binomial(n, θ)
- $H_0 : \theta = \frac{1}{2}$ vs $H_1 : \theta \neq \frac{1}{2}$
- Ask to keep $\mathbb{P}(\textit{reject } H_0 \mid H_0) \leq 5\%$
- Ask for $\mathbb{P}(\textit{reject } H_0 \mid |\theta - \frac{1}{2}| \geq 0.1) \geq 92\%$
- Assume $\bar{X} \sim N(\theta, \theta/4\sqrt{n})$ in the range $\theta \in [0.4, 0.6]$
- Need n such that
  $[F^{-1}_{N(0,1)}(1 - 5\%/2) + F^{-1}_{N(0,1)}(1 - 8\%)] \times s.d. \geq 0.1$

Confidence Intervals

- Parameter $\theta \in \mathbb{R}$
- 95% CI: Statistics L, R such that

$$\forall \theta,\ \mathbb{P}_\theta[L \leq \theta \leq R] \geq 95\%$$

- Not unique—one sided or two sided, etc
- Complement of the critical regions for testing $H_0 : \theta = \hat{\theta}$, i.e. that the MLE is the right parameter.
- N.B. Here θ is fixed and the statistics L, R are random.

Normal confidence intervals

- $X_1, \ldots, X_n \sim N(\theta, 1)$ iidrv
- MLE $\bar{X} = \hat{\theta} \sim N(\theta, 1/n)$
- $$\mathbb{P}[\theta - \frac{1.96}{\sqrt{n}} \leq \hat{\theta} \leq \theta + \frac{1.96}{\sqrt{n}}) = \mathbb{P}[\theta \in (\hat{\theta} - \frac{1.96}{\sqrt{n}}, \hat{\theta} + \frac{1.96}{\sqrt{n}})] = 95\%$$
- $$\mathbb{P}[\hat{\theta} \geq \theta - \frac{1.64}{\sqrt{n}}) = \mathbb{P}[\theta \in (-\infty, \hat{\theta} + \frac{1.64}{\sqrt{n}})] = 95\%$$
- $$\mathbb{P}[\hat{\theta} \leq \theta + \frac{1.64}{\sqrt{n}}) = \mathbb{P}[\theta \in (\hat{\theta} - \frac{1.64}{\sqrt{n}}, \infty)] = 95\%$$

t confidence intervals

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ iidrv
- Sample mean $\bar{X} = \hat{\mu} \sim N(\theta, 1/n)$, sample variance S^2
- $$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$
- Choose q such that
  $F_{t_{n-1}}(q) - F_{t_{n-1}}(-q) = \mathbb{P}(-q \leq A \leq q \mid A \sim t_{n-1}) = 95\%$
- Hypothesis test: Under $H_0 : \mu = \mu_0$,
  $F_{t_{n-1}}\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right) \sim \mathit{Uniform}(0,1)$
- Confidence interval: $\forall \mu$, $\mathbb{P}(\mu \in (\bar{X} - q\frac{S}{\sqrt{n}}, \bar{X} + q\frac{S}{\sqrt{n}})) = 95\%$

Hypothesis test for contingency table

- $m \times n$ contingency table
- H_0 :properties are independent $p_{i,j} = a_i \times b_j$;
  $\sum_i a_i = 1$, $\sum_j b_j = 1$ so $m + n - 2$ degrees of freedom
- H_1 :properties are not independent.
  $m \times n - 1$ degrees of freedom
- Number of observations $N = \sum_{i,j} O_{i,j}$
- Expected number of observation under H_0 is
  $E_{i,j} := N \times (\sum O_{i,k}/N) \times (\sum O_{k,j}/N)$
- Asymptotically under H_0 by Wilk's theorem:

$$-2 \log \frac{\prod (E_{i,j}/N)^{O_{i,j}}}{\prod (O_{i,j}/N)^{O_{i,j}}} \approx \chi^2_{(m-1)(n-1)}$$

- Pearson's χ^2 test statistic $\sum (O_{i,j} - E_{i,j})^2 / E_i$ approximates above if $\min_i E_i \geq 5$.
- Large values: reject independence. Small values: faked data?

Variance stabilizing transforms

- $X \sim f(\cdot \mid \theta)$ and $\mathbb{E}X = \theta$
- $\mathrm{Var}_\theta(X) =: V(\theta)$
- $Y := X/\sqrt{V(\theta)}$ has variance 1
- Taylor's theorem:
  $\mathrm{Var}(g(X)) \approx \mathrm{Var}(\theta + g'(\theta) Y \sqrt{V(\theta)}) \approx g'(\theta)^2 V(\theta)$
- Want to find g such that $\mathrm{Var}(g(X)) \approx g'(\theta)^2 V(\theta) \approx$constant

$$g'(\theta) \approx \mathit{constant} \times V(\theta)^{-1/2}$$

$$g(\theta) = \int^\theta V(u)^{-1/2} du$$

- Poisson: $V(\theta) = \theta \to g(X) = \sqrt{X}$
- Exponential mean θ: $V(\theta) = \theta^2 \to g(X) = \log X$

Bayesian statistics

- Parameter θ with prior belief $f(\theta)$
- Data $X \sim f(X \mid \theta)$
- Joint distribution $f(X, \theta) = f(\theta) f(X \mid \theta)$, $\int_\theta \int_x f(x, \theta) dx d\theta = 1$
- Bayes theorem, Bayes' theorem, Bayes's theorem:

$$f(\theta \mid x) = \frac{f(x, \theta)}{\int_t f(x, t) dt} = \frac{f(\theta) f(x \mid \theta)}{Z(x)}$$

i.e. "Posterior is proportional to prior times likelihood"
Can generally ignore the normalizing constant $Z(x)$

Bayesian statistics

- Instead of MLE $\hat{\theta}$, we can look at properties of the posterior distribution
  - $\delta =$"posterior mean" minimizes the expected square error
  - $\delta =$"posterior median" minimes the expected absolute error
- The prior distribution does not need to be a real probability distribution.
  If $\int f(\theta)d\theta = \infty$ call it an improper prior.
- For a random sample of size n, as $n \to \infty$, the prior becomes less important.
  Asymptotically $f(\theta \mid X_1, \ldots, X_n) \sim N(\theta, I(\theta)^{-1}/n)$ (just like MLE).
- The "exception" to this rule is if the prior is way off, i.e. taking $\theta \sim N(0,1)$ or $\theta \sim Uniform(0,1)$ when θ is really 100

Credible intervals

- For Bayesian, credible intervals replace confidence intervals.
- A 95% credible interval is an interval covering 95% of the posterior

$$\int_{L(x)}^{R(x)} f(\theta \mid x) d\theta = 95\% \leftrightarrow \mathbb{P}_{posterior}(\theta \in (L(x), R(x))) = 95\%$$

- Unlike the frequentist case, given the data, "$\theta \in (L(x), R(x))$?" is still officially random. ☺

Where do priors come from?

- Non-informative priors: make up something so broad that it is guarenteed to cover all but the most unrealistic values of θ.
- OR: ask an expert
- Conjugate priors: some pairs
  - normal prior and normal likelihood
  - beta prior and binomial likelihood
  - beta prior and geometric likelihood
  - gamma prior and poisson likelihood
  - gamma prior and normal likelihood
  - gamma prior and gamma likelihood
  - etc

  work out nicely analytically, so are often used.
- Jeffrey's prior $f(\theta) \propto \sqrt{I(\theta)}$ invariant under reparametrization
- If the prior looks a lot like the posterior, your experiment is rather questionable.

Jeffrey's prior example

- Likelihood Bernoulli(θ)
- Could call the Uniform$(0, 1)$ distribution an uninformative prior
- Jeffrey's prior:

$$f(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta} \log f(x \mid \theta)\right) = Var\left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right)$$

$$= Var\left(\frac{X-\theta}{\theta(1-\theta)}\right) = \frac{1}{\theta(1-\theta)}$$

$$f(x \mid \theta) \propto I(\theta)^{-1} \rightarrow f(x \mid \theta) = Beta\left(\frac{1}{2}, \frac{1}{2}\right)$$

- Observe n samples: k 1s and $n-k$ 0s
  Posterior $= Beta(\frac{1}{2} + k, \frac{1}{2} + n - k)$
- Credible interval: 2.5% and 97.5% quantiles of the posterior distribution [qbeta]