Statistics Review

John Semerdjian

Basics

Probability Mass Functions (PMF): a function that gives the probability that a discrete random variable is exactly equal to some value. PMF is normalized so total probability is 1.

Probability Density Function (PDF): derivative of CDF. Probability density measures probability per unit of x. In order to get a probability mass, you must integrate over x.

Cumulative Distribution Functions (CDF): the function that maps from a value to its percentile rank.

Effect Size

Measure of the strength of a phenomenon. Complements statistical hypothesis testing, and play an important role in statistical power analyses, sample size planning. Other examples of ES: correlation, regression coefficient, odds ratios, mean difference, risk.

Cohen's d: difference between groups to the variability within groups (σ is pooled):

$$\text{Cohen's } d = \frac{\bar{x_1} - \bar{x_2}}{\sigma}$$

Odds Ratio:

$$\text{Odds Ratio} = \frac{p}{p-1} \leftarrow \text{not symmetric}$$
$$\text{Log Odds (aka Logit)} = \log\left(\frac{p}{p-1}\right) \leftarrow \text{symmetric}$$

Distributions

Normal

$$\text{PDF} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
$$\underbrace{\text{Log Likelihood}}_{\text{for } n \text{ independent } N(\mu,1)} = \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - n\log(\sqrt{2\pi})$$
$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$$
$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Binomial

$$\text{PMF, Binomial Formula, } P(k \mid n,p) = \binom{n}{k} p^k (1-p)^{n-k} \leftarrow \text{Sum of Binomials}$$

$$\begin{aligned}
\text{Binomial}(1,p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}, \text{where } x_i = 0 \text{ or } 1 \\
&= \underbrace{S \log p + (n-S)\log(1-p)}_{\text{Let } S = x_1 + \ldots + x_n} \\
\text{Log Likelihood} &= \sum_{i=1}^{n} x_i \log p + (1-x_i)\log(1-p) \\
\mu &= np \\
\sigma^2 &= np(1-p) \\
\text{SE for number} &= \sqrt{np(1-p)} \\
\text{SE for percent} &= \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

Negative Binomial

Probability of that it takes n flips to get k heads

$$\text{PMF, Negative Binomial Formula, } P(n \mid k,p) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Exponential

Come up when we look at a series of events and measure the times between events, called interarrival times. If the events are equally likely to occur at any time, the distribution of interarrival times tends to look like an exponential distribution.

$$\begin{aligned}
\text{CDF}(x) &= 1 - e^{-\lambda x} \\
\text{PDF}(x) &= \lambda e^{-\lambda x}
\end{aligned}$$

- λ: can be interpreted as a rate; that is, the number of events that occur, on average, in a unit of time
- $\frac{1}{\lambda}$ is the mean of an exponential distribution.

Beta

The Beta distribution can be understood as representing a probability distribution of probabilities. It represents all the possible values of a probability when we don't know what that probability is.

Is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions.

Method of Moments: Use MLE to fit probability distribution, or use method of moments and solve for α and β:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

A/B Testing

Frequentist A/B

with equal sample sizes

$$\text{SE}_{\text{diff. in prop.}} = \sqrt{\frac{p_A(1-p_A)}{N_A} + \frac{p_B(1-p_B)}{N_B}}$$

with pooled SE

$$\begin{aligned}
p_{\text{diff}} &= \frac{n_B}{N_B} - \frac{n_A}{N_A} \\
p_{\text{pooled}} &= \frac{n_B + n_A}{N_B + N_A} \\
\text{SE}_{\text{pooled}} &= \sqrt{p_{\text{pooled}} \times (1 - p_{\text{pooled}}) \times \left(\frac{1}{N_B} + \frac{1}{N_A}\right)} \\
\text{CI} &= p_{\text{diff}} \pm 1.96 \times \text{SE}_{\text{pooled}}
\end{aligned}$$

Bayesian A/B

Calculate probability that the Beta distribution of A is greater than B by simulating draws from the posterior distribution.

1. Calcuate priors for each group, e.g. $\text{Beta}(\alpha_0 = 1, \beta_0 = 1)$
2. Run experiment
3. Update priors for each group using data, e.g. $\text{Beta}(\alpha_0 + \text{successes}, \beta_0 + \text{failures})$
4. Simulate draws from posterior distribution

```r
# priors
alpha_0 = 1; beta_0 = 1

# run 100 trials per group
success_a = 42; total_a = 100
success_b = 47; total_b = 100

# simulate 10,000 draws from posterior
A = rbeta(10000, alpha_0 + success_a, beta_0  + total_a - success_a)
B = rbeta(10000, alpha_0 + success_b, beta_0  + total_b - success_b)

# probability that B is larger than A
mean(B > A)

# 95% credible intervals
quantile(B - A, c(0.25, 0.975))
```

Sample Size Calculations

Compare 2 Proportions

We want to choose N such that 80% of the possible 95% confidence interval will not include 0.5. When N increases, the estimate becomes closer to the true value, and the width of the confidence interval decreases.

$$p_A = 0.5,\ P_B = 0.6$$
$$z_{1-\alpha/2} = 1.96, \text{ where } \alpha = 0.05$$
$$z_{1-\beta} = 0.84, \text{ where } \beta = 0.2$$

Solve for N where:

$$\underbrace{0.5}_{p_A} + 1.96 \times \text{SE} = \underbrace{0.6}_{p_B} - \underbrace{0.84}_{\text{power}} \times \text{SE}$$

$$\text{SE} = \frac{0.5}{\sqrt{N}}$$

$$196 = (2/8 \times 0.49/0.1)^2$$

To have 80% power, the true value of the parameter must be 2.8 standard errors from the comparison point.

Alternative formula

$$N \geq \big(p_A(1-p_A) + p_B(1-p_B)\big)\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{p_A - p_B}\right)^2 = 385 \text{ per group}$$

```
p0 = 0.5; p1 = 0.6
alpha = 0.05; power = 0.8
z_alpha = qnorm(1-alpha/2)
z_power = qnorm(power)

# sample size per group
(p0*(1-p0) + p1*(1-p1))*((z_alpha + z_power)/(p1-p0))^2
```

Relationships between variables

Covariance: a measure of the tendency of two variables to vary together.

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Pearson correlation: works well if the relationship between variables is linear and if the variables are roughly normal, but is not robust to outliers. $\rho \in [-1, 1]$.

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

Spearman rank correlation: transforms values into ranks. If x is skewed and/or has outliers, use Spearman instead of Pearson.

Confidence Intervals

Definition: If we were to draw 100 samples from same population, approximately 95 of them would contain the parameter.

Standard Error: measure of variability due to sampling error. Different samples drawn from that same population would in general have different values of the sample mean, so there is a distribution of sampled means.

When estimating percentages, it is the absolute size of the sample which determines accuracy, not the size relative to the population.

Highest Density Interval (HDI) / Credible Interval: Bayesian version of confidence interval. From the posterior distribution we can get the average and the credible interval (i.e. the uncertainty). If a parameter value lies within the 95% credible interval, it is said to be among the credible values.

Region of Practical Equivalence (ROPE): Bayesian version of practical significance. A small range of values that are considered to be practically equivalent to the null value for purposes of the particular application. A parameter value is declared not credible if its entire ROPE lies outside the 95% HDI of the posterior distribution of that parameter.

Normal Approximation: Use normal distribution to approximate the distribution of error about a binomially-distributed observation. The Central Limit Theorem applies poorly to this distribution with a sample size less than 30 or where the proportion is close to 0 or 1.

$$\text{CI} = p \pm 1.96 \times \underbrace{\text{SE for percent}}_{\text{w/ replacement}}$$

Wilson Interval: An improvement over the normal approximation interval. (`binom.test` uses Clopper-Pearson interval.)

Frequentist vs. Bayesian

	Bayesian	Frequentist
Data	fixed	random
Parameter	random	fixed, but unknown

Hypothesis Testing

Parametric

z Test: compare random sample of measuresments with a large parent group whose mean and standard deviation are known.

$$z = \frac{\mu^{pop.} - \mu_i}{\text{SE}} = \frac{\mu^{pop.} - \mu_i}{\sigma^{pop.}/\sqrt{n}}$$

Two sample z Test:

$$z = \frac{\text{observed diff.} - \text{expected diff.}}{\text{SE for diff.}} = \frac{\text{observed diff.} - 0}{\sqrt{x_1^2 + x_2^2}}$$

Student's t-Test: compare a random sample with a large parent group whose mean is known, but whose standard deviation is not known. Null: there is no difference between the mean of the sample group and the mean of the large parent group.

$$t = \frac{\mu^{pop.} - \mu_i}{\text{SE}} = \frac{\mu^{pop.} - \mu_i}{\sigma^{sample}/\sqrt{n}}$$

Welch's t-Test: a variant of the classic Student's t-test which allows for two samples of different size, and possibly different variance. Unlike in Student's t-test, the denominator is not based on a pooled variance estimate.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\text{d.f.} = \frac{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^2}{(\sigma_1^2/\sigma_1)^2/(n_1 - 1) + (\sigma_2^2/n_2)^2/(n_2 - 1)}$$

ANOVA: test of whether or not the means of several groups are equal; generalizes the t-test to > 2 groups.

Chi-Squared Test: compare any 3 or more groups by observed and expected frequencies. In all cases the expected number should be 5 or more in each cell.

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$\text{d.f.} = \text{number of terms in } \chi^2 - 1$$

Yates' Chi-Squared Test: modified version of Chi-Squared used on 2x2 contingency tables for comparing:

1. 2 random **binomial** samples for evidence of difference between samples
2. matched observations made on a single random sample group, for evidence of association between 2 qualities

Expected number in each cell should be > 5. For smaller values use Fisher's Test

Fisher's Test: tests null hypothesis of independence of rows and columns in a contingency table with fixed marginals. Same comparisons as Yate's Chi-Squared Test. Test is based on the discovery that the exact probability of getting any particular values within cells by chance is given by the hypergeometric formula:

$$P = \frac{n_1! \times n_2! \times n_3! \times n_4!}{N! \times a! \times b! \times c! \times d!}$$

Non-Parametric

Permutation test: model the null hypothesis by shuffling data from both samples, and create two new groups. Observe how often your test statistic (e.g., difference between means, correlation, sum of the absolute differences) is as extreme as your original estimate (i.e., p-value). Can perform one-sided (use absolute values) or two-sided versions.

Wilcoxon Rank Sum Test (Mann–Whitney Test): compare 2 unmatched random samples of measurements. Greater efficiency than the t-test on non-normal distributions, such as a mixture of normal distributions, and is nearly as efficient as t-test on normal distributions.

Wilcoxon Signed Rank Test (similar to paired t-test): compare 2 random samples of measurements which are matched. Assess whether population mean ranks differ. It can be used when population cannot be assumed to be normally distributed.

Kruskal-Wallis Rank Sum Test (expands Wilcoxon Rank Sum to $>$ 2 groups): compare 3 or more unmatched random samples of measurements.

Multiple Hypothesis Testing

Frequentist

Bonferroni correction: Controls the probability of at least one Type I error. Bonferroni (and other FWER corrections) are too conservative.

Benjimini Hochberg: FDR-controlling procedures are designed to control the expected proportion of rejected null hypotheses that were incorrect rejections ("false discoveries"). FDR-controlling procedures provide less stringent control of Type I errors compared to FWER controlling procedures. Thus, FDR-controlling procedures have greater power, at the cost of increased rates of Type I errors.

1. Sort and rank p values, 1, ..., N
2. Select an FDR rate, say 0.05
3. Adjusted P value $= \dfrac{\text{Rank}}{N} \times 0.05$
4. If the adjusted P value is smaller than the false discovery rate, the test is significant

Bayesian

In Bayesian analysis the interpretation of the data is not influenced by the experimenter's intentions. A Bayesian analysis yields a posterior distribution over the parameters of the model. The posterior is the complete implication of the data.

Bayesian analysis eschews the use of p values as a criterion for decision making because the false positive rate depends on the experimenter's intentions.

Errors

Type I (False Positive Rate), α: we reject the null hypothesis but we shouldn't have

Type II (False Negative Rate), β: we don't reject the null, but we should have

Power (aka "Sensitivity"), $1 - \beta$: the probability of detecting a true effect

$$\text{Power} = P(\text{reject } H_0 \mid H_1 \text{is true})$$

- As a rule of thumb, a power of 80% is considered acceptable.
- As the power increases, the chances of a Type II error decrease.
- Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size.

Causal Inference

Selection Bias: participants select themselves into different treatments

Confounding: variables that effect both treatment and outcome that are omitted during comparisons

Internal Validity: a study in which casual inferences are merited for a specific sample or population

External Validity: a study in which casual inferences can be generalized to a broader populaiton of interest

Intention to Treat:

Cross-over:

Neyman Box Model

Types of Casual Inferences

Randomized Experiments

We can't estimate individual causal effects, but we can deisgn studies to estimate the population average treatment effect

Regression

Only works if all confounding variables are included and if the model is correctly specified. We estimate the average effect on y for each additional unit of T.

Stable unit treatment value assumption (SUTVA): Assumption (for casual effects) that the treatment assignment for one individual (unit) does not affect the outcome of another.

Interactions: modeling interactions is important when we care about differences in the treatment effect for different groups, and poststratification then arises naturally if a population average estimate is of interest.

Poststratification: the analysis of an unstratified sample, breaking the data into strata and reweighting as would have been done had the survey actually been stratified. Stratification can adjust for potential differences between sample and population using the survey design; posstratification makes such adjustments in the data analysis.

Ignorability of Treatment Assignment (Observational studies): Units are randomly assigned to treatment conditions on the confounding covariates, even though no actual randomized assigned took place. We expect any two classes at the same levels of the confounding covariates to have had the sample probability of receiving the supplemental version of the treatment. If ignorability holds, then casual inferences can be made without modeling the treatment assignment process. In general, one can never prove that the treatment assignment process in an observational study is ignorable.

$$y^0, y^1 \perp T \mid X$$

Lack of Complete Overlap and Imbalance: covariate distributions should be the same across treatment groups.

IVLS

Propensity Score Matching

Matching