Statistics & Experimental Design with R

Barbara Kitchenham

Keele University

Basic Statistical Theory

Part 2

Probability Distributions

- Frequency function
  - Also called probability density function for continuous variables
  - Integral referred to as “cumulative distribution function”
- Three properties:

$$f(x) \geq 0$$

$$F(x) = \int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{a}^{b} f(x)dx = P\{a < x < b\}$$

Normal (Gaussian) Distribution

- Probability distribution $x \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

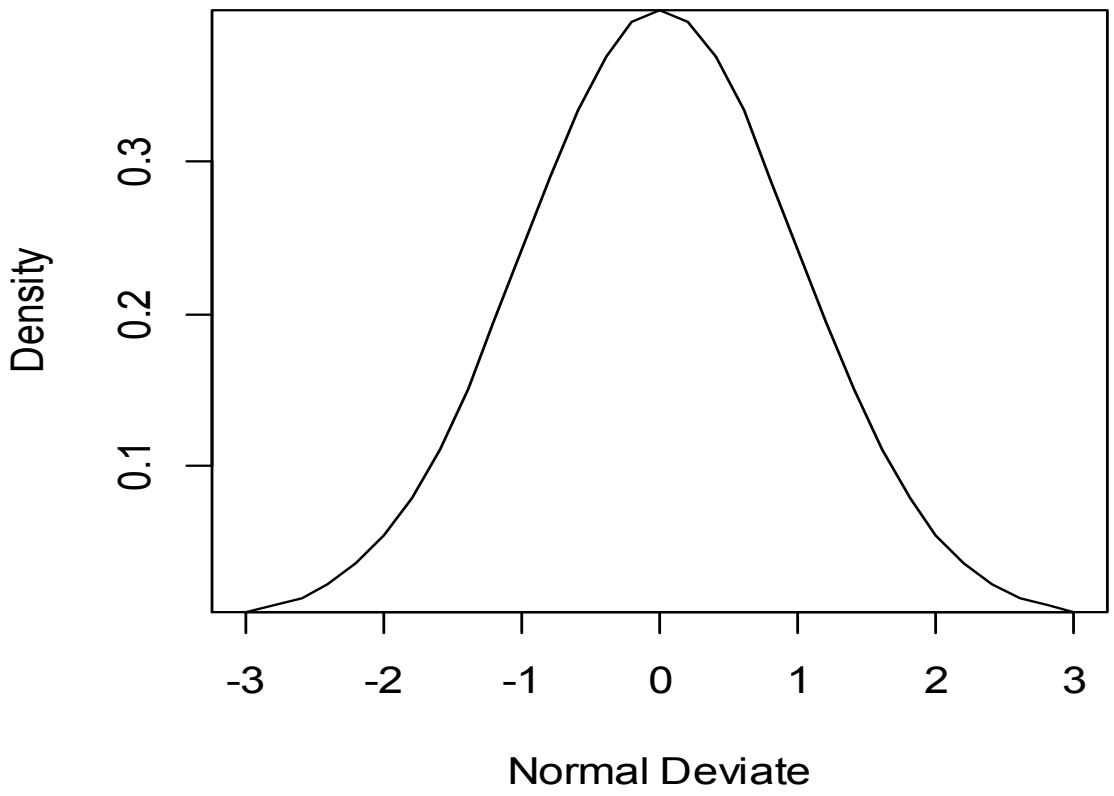
- Any normal distribution can be *standardized*, $z \sim N(0,1)$ letting $z = \frac{x-\mu}{\sigma}$

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- Always symmetric about mean (μ)
  - $P\{-\sigma < x < \sigma) \sim 0.68$
  - $P\{-2\sigma < x < 2\sigma) \sim 0.95$

Keele University

Normal distribution



Moments

- Moments – a measure of the shape of a set of points
  - Moments about origin

$$\mu'_k = \int_{-\infty}^{\infty} x^k f(x)dx \qquad \mu = \int_{-\infty}^{\infty} x f(x)dx$$

  - Moments about mean

$$\mu_k = \int_{-\infty}^{\infty} (x - \mu'_1)^2 f(x)dx \qquad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

  - μ & σ^2 define the Normal distribution
  - Third (& odd>3) moments about mean (skewness) =0
  - Fourth moment about mean (kurstosis)=3

Expectations – Fuctions of Variables

- Expected value of a function *h(x)* of random variable *x* is defined as:

$$E[h(x)] = \int_{-\infty}^{\infty} h(x) f(x) dx$$

  - Provide a precise definition of important quantities
  - Provide link between samples and populations
- If $h(x)=x$, $E[x]= \mu_x$
- Arithmetic transformations of functions of random variables easy to handle
  - $E[b+cx]= b+c\mu_x$
  - $E[x_1+x_2+x_3+\ldots]= \Sigma\mu_i$

Expectations of Variance

- Expected value of var x =

$$E[var(x)] = \int_{-\infty}^{\infty}(x-\mu_x)^2 f(x)dx$$

- For the sum or difference of two variables

$$E[var(x+y)] = \sigma_x^2 + \sigma_y^2 + 2\sigma_{x,y}$$
$$E[var(x-y)] = \sigma_x^2 + \sigma_y^2 - 2\sigma_{x,y}$$

  - If x and y are independent $2\sigma_{x,y} = 0$

- Arithmetic transformations are allowed

$$E[var(bx+c)] = b^2\sigma_x^2$$

Properties of Normal Variables

- If$\{X_1,\ldots,X_n\}$ are a set of independent, identically distributed Normal variables of size *n*
  - Each with mean=μ and variance σ^2
  - E[mean= $\Sigma X_i/n$]= μ
  - E[var $(\Sigma X_i/n)$]= $(\Sigma\sigma^2)/n^2 = \sigma^2/n$
  - **$\Sigma X_i/n$ is ~N(μ,σ^2/n)**
- Variance of $\{X_1,\ldots,X_n\}$ is chi-squared distribution with *n* degrees of freedom
  - $\Sigma(X_i- \mu)^2/n \sim\sigma^2\chi^2_n/n$
  - Expected value of χ^2_n=n, var(χ^2_n)=2n
  - Var$(\Sigma(X_i- \mu)^2/n)= 2n\sigma^4/n^2=2\sigma^4/n$

Maximum Likelihood -1

- Generic method of estimating parameters of a distribution
- Likelihood function (L)
  - Joint distribution of elements in a sample given the values of a parameter θ

$$L = Prob(x_1, x_2, \ldots x_n \backslash \theta)$$

  - Parameter estimated by
    - Differentiating L (usually Log(L)) with respect to θ,
    - Equating equation derivatives to zero
    - Solving equations
    - Accept solution for which second derivative is negative

Maximum Likelihood -2

$$L = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x_1-\mu)/\sigma^2]} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x_2-\mu)/\sigma^2]} \cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x_n-\mu)/\sigma^2]}$$

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}[\sum_{i=1}^{n}(x_i-\mu)/\sigma^2]} \qquad Log(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$\frac{\partial LogL}{\partial \mu} = 0 = \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma^2}\right) \qquad \frac{\partial LogL}{\partial \sigma^2} = 0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

- L is like Bayesian model with no Prior
- ME estimate of sigma is biased
- When f(x) Normal, Log(L) is chi-squared with n degrees of freedom
- Log(L) is used in many statistical tests

Importance of Normal Distribution

- Law of large numbers
  - The *average* of the results obtained from a number of "trials"
    - Should be close to expected value
    - Becomes closers as more trials are performed
- Central limit theorem
  - If$\{X_1,\ldots,X_n\}$ are a set of independent, identically distributed variables of size *n*
  - $S_n = \Sigma X_i/n$ is approximately $\sim N(\mu,\sigma^2/n)$
  - **Irrespective of distribution of X's**
    - **Assuming finite X_i have variances**
- Normal distribution assumed to occur as the sum of many small independent effects

Implications

- Classical methods
  - With large enough sample size, can assume the *mean* of a sample is Normally distributed
    - Can use properties of Normal distribution
      - E.g. Standard unit distribution can be used to construct confidence intervals
  - An immense body of statistical methods available if parameters/data are normal
  - Many guidelines for transforming the data to increase Normality

Normal approximations

- Binomial Distribution
- Probability of *x* successes in *n* trials

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

  - *p* is probability of success for a specific trial
  - Expected value of *p* is $\hat{p} = x/n$
  - Expected variance of *p is* $s_p^2 = \frac{\hat{p}(1-\hat{p})}{n} = \hat{p}\hat{q}/n$
- Approximately Normal
  - If *n* large (>30)
  - *p* not too far from 0.5
  - Confidence intervals for *x* or *p* based on Normal distribution
  - With “corrections” for discrete distribution

Confidence Limits of Mean

- Assume random sample
- Mean is approximately Normal $\bar{x} = \sum_i^n x_i / n$

$$s^2 = \sum_i^n (x_i - \bar{x})^2/(n-1) \qquad var(\bar{x}) = \frac{s^2}{n}$$

  - For 95% confidence intervals

$$\int_{-a}^{a} f(x)dx = P\{-a < \bar{x} < a\} = 95\%$$

  - For unit normal deviate

$$\int_{-a}^{a} f(x)dx = P\{-a < x < a\} = 95\% \qquad -a = .025\ quartile = -1.96$$

  - For random sample, confidence limit of mean

$$CL = \bar{x} \mp 1.96 \times s/n$$

Confidence Limits of Differences

- Independent random samples from two groups, want to investigate $\bar{x}_1 - \bar{x}_2$

$$var(\bar{x}_1 - \bar{x}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- Assuming variance same in each group

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

$$var(\bar{x}_1 - \bar{x}_2) = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Student's *t* Distribution

- Provide means of correcting for small samples
  - When estimates are less reliable (e.g. <30 per group)

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

  - Degrees of freedom = *n-1*
  - Confidence limits found as usual (assuming α level)

$$CL = \bar{x} \mp t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

Approximations &Transformations

- Pearson correlation coefficient
- Association between two variables (x,y) (measured on same item)

$$\rho = \frac{covariance(x,y)}{\sigma_x \sigma_y} \qquad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- For large n>100 $s.e\ of\ r = \frac{1 - r^2}{\sqrt{n}}$
- For small n, use Normal transformation

$$z = \ln\left(\frac{1+r}{1-r}\right) \quad var(z) = 1/(n-3)$$

Problem

- How large a sample is needed for good Normal approximation?
  - 30+? Point where "t" distribution and Normal distribution converge
- Systematic studies of Non-normality
  - "Heavy" tails (i.e. many outliers) but symmetric
  - Skewed but "light-tailed"
  - Heavy-tailed and skewed
- Show classical methods more vulnerable than expected
  - For skewed distributions the mean may be far from "typical"
  - Heavy-tails increase the variance
    - Making it possible to miss true effects
    - Also tests for non-Normality have **low power**
      - They are vulnerable to Type 2 Errors

Keele University

The Workshop Approach

- We have reviewed some important classic techniques
- But
  - Will continue to concentrate on conventional approaches
  - But will introduce some new approaches
    - Particularly ones that let you visualise your data
  - Review some recent approaches to robust analysis
- However from now approaches will be illustrated with SE data