Lecture 4: Probability

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

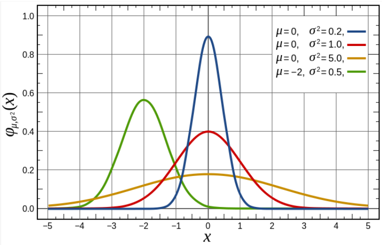
10/18/2016

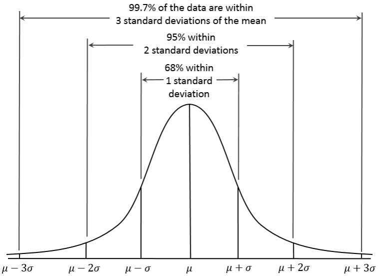
Overview

Gaussian Distribution

Also called the Normal distribution. Due to the central limit theorem, this distribution is very common in statistics.

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





Left: Wikipedia, Right: By Dan Kernler - Own work, CC BY-SA 4.0

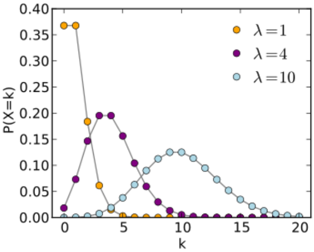
Common Probability Distributions: Poisson Distribution

Poisson Distribution

The Poisson distribution describes the likelihood of an event occurring in a fixed interval of time if the average event rate (λ) is known.

$$P(\text{observe k events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

k is a non-negative integer. Mean is $\mu = \lambda$ and standard deviation is $\sigma = \sqrt{\lambda}$



Common Probability Distributions: Binomial Distribution

Bernoulli Random Variable

- A Bernoulli random variable has two possible outcomes "success" (1) or failure (0).
- If X is a random variable with $P(X = 1) = p$ and $P(X = 0) = 1 - p$, then X is a Bernoulli random variable with mean $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.

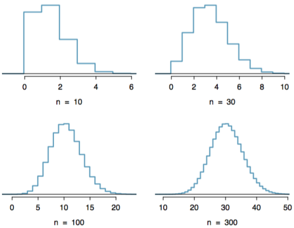
Let Y denote the number of successes in the first n trials, then the probability distribution of Y is the **binomial distribution**:

$$P(y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{k!(n-k)!} p^y (1-p)^{n-y}$$

Binomial Distribution

Normal Approximation to the Binomial Distribution

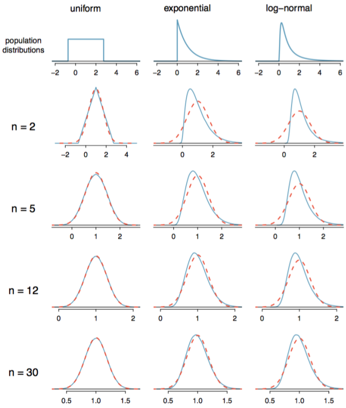
If the number of trials n is sufficiently large, then the binomial approximation is approximately equal to the normal distribution with mean $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. The condition is that $np > 10$ and $n(1-p) > 10$.



Binomial distribution with $p = 0.10$, n shown below histogram. [Diez, 2016]

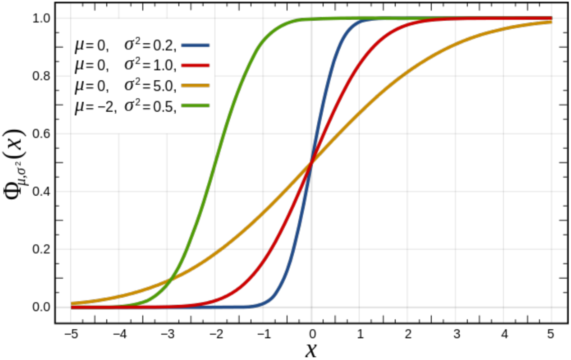
Central Limit Theorem

The mean of a large number of independent, identically distributed variables will be approximately normal, for all underlying distributions.



Graphic from [Diez, 2016]

Cumulative Distribution Function



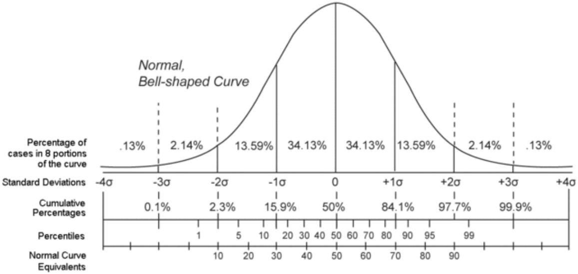
https://commons.wikimedia.org/w/index.php?curid=3817960By Inductiveload - self-made, Mathematica, Inkscape, Public Domain

Quantiles

Quantile

Quantiles are cut-points that divide a probability distribution into contiguous intervals with equal probabilities. Special cases:

- 2-quantile is the median
- 4-quantiles are *quartiles*; the difference between the upper and lower quartiles is the *interquartile range*
- 100-quantiles are percentiles



Standardizing

Z-score

The *Z-score* of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

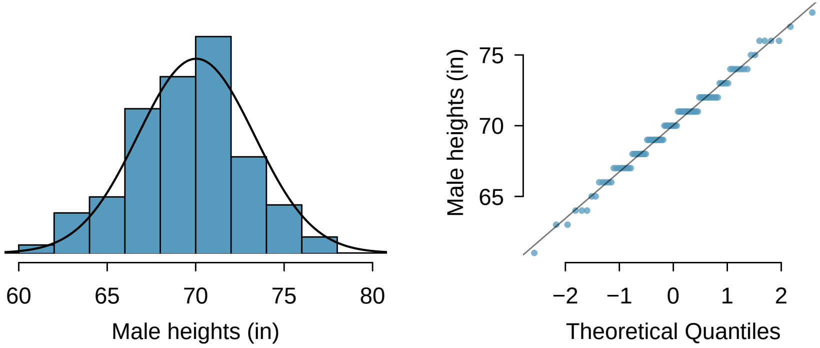
- Z-scores are defined for distributions of any shape, but with the normal distribution we can use them to calculate percentiles.
- Z-score $\rightarrow$ percentile, use scipy.stats.norm.cdf(z-score)
- percentile $\rightarrow$ z-score, use scipy.stats.norm.ppf(percentile)
- Observations that are more than two standard deviations away from the mean are unusual ($2 \times$ norm.cdf(-2) $= 0.045$)

Evaluating the normal approximation

Slides from this section, noted with (OI) in the title, are copied from OpenIntro Chapter 3: Distributions of Random Variables.
Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Normal probability plot (OI)

A histogram and *normal probability plot* of a sample of 100 male heights.

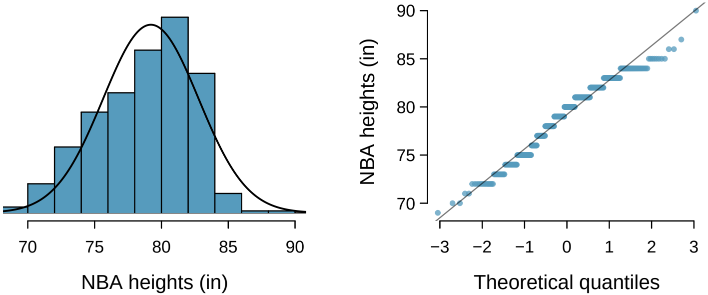


Anatomy of a normal probability plot (OI)

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

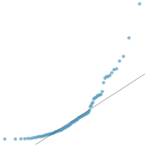
Normal probability plot (OI)

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

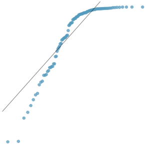


Why do the points on the normal probability have jumps?

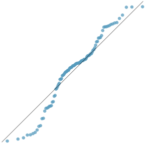
Normal probability plot and skewness (OI)



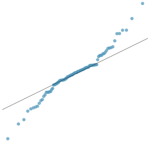
Right skew - Points bend up and to the left of the line.



Left skew- Points bend down and to the right of the line.



Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.



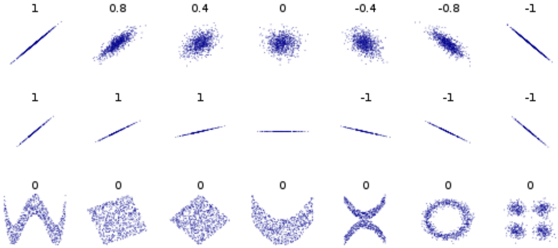
Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

Correlation

Correlation Coefficient

Also known as Pearson's [product-moment] coefficient measures the linear correlation between two random variables X and Y.

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$



Bayes' Theorem

Bayes' theorem provides a method to calculate the probability of an event (A) in a certain context (B), based on knowing the overall probability of the event, the overall probability of the context, and the probability of the context given the event:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of Total Probability

Bayes' Theorem can be derived from the Law of Total Probability:

$$P(E) = \sum_i g(x_i)P(E|X = x_i)$$

where g(x) is the probability distribution for x.

References

Kyle Siegrist
Probability, Mathematical Statistics, Stochastic Processes

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 2-3
Data Science from Scratch, Chapter 6

For discussion
Stan - The Bayesian Data Scientist's Best Friend
For next week
OpenIntro Statistics, Chapters 4-6

Exercise

Lesson4_Distributions.ipynb

- Assess the normality of a data set