Statistical Inference (WIP)

Albert Mata

18/04/2017

Contents

1. Introduction

We'll define statistical inference as the process of generating conclusions about a population from a noisy sample. Without statistical inference we're simply living within our data. With statistical inference, we're trying to generate new knowledge. Statistical inference is about describing populations using data.

Statistical inference is about analyzing data and drawing conclusions about it, usually so we can fit a mathematical model to the data (a probability model or some type of predictive model, like a regression mode). The value derived from fitting a useful model is often the payoff of laborious experimentation and data collection.

There are so many shades of gray between the styles of inferences that it is hard to pin down most modern statisticians as either Bayesian or frequentist. In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences. This is the most popular style of inference by far.

More mathematical content on these YouTube playlists:

- *Math Biostat Boot Camp 1*
- *Math Biostat Boot Camp 2*

Combinatorics

Permutations: arrangement of r out of n distinct objects (order is important):

$$P(n,r) = \frac{n!}{(n-r)!}$$

Combinations: selecting r out of n distinct objects(order is not important):

$$C(n,r) = \frac{n!}{r!(n-r)!}$$

```
# C(n,r) = choose(n, r)
choose(3, 2)
## [1] 3
```

```
# Factorials can be calculated using gamma function knowing that n! = gamma(n+1)
gamma(5)
## [1] 24
```

Expected values

Most usual expected values for population distributions are:

- **Mean**: characterizes the center of a density or mass function. $E[X]$ or μ for population. $\bar{X}$ for sample. $E[X] = \sum_x xp(x)$
- **Variance**: characterizes how spread out a density is. $Var(X)$ or σ^2 for population. S^2 for sample. $Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$
- **Skewness**: considers how much a density is pulled toward high or low values.

The average of ten randomly sampled people's height is itself a random variable, in the same way that the average of ten die rolls is itself a random number. Thus, the distribution of heights gives rise to the distribution of averages of ten heights in the same way that distribution associated with a die roll gives rise to the distribution of the average of ten dice. And what does the distribution of averages look like? It's centered at the same spot as the original distribution! Thus, the distribution of the estimator (the sample mean) is centered at the distribution of what it's estimating (the population mean). **When the expected value of an estimator is what its trying to estimate, we say that the estimator is unbiased**.

IID random variables

Random variables are said to be independent and identically distributed (*iid*) if they are independent and all are drawn from the same population.

2. Probability

Probability for us will be the long run proportion of times some occurs in repeated unrelated realizations.

Every random variable has an associated probability distribution function. This function is called a probability mass function in the case of a discrete random variable or probability density function in the case of a continuous random variable.

Densities and mass functions for random variables are the best starting point for modeling and thinking about probabilities for numeric outcomes of experiments:

- A probability density function (PDF), or density of a **continuous random variable**, is a function, whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample. The PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value (which would be 0, since there are an infinite set of possible values to begin with).
- A probability mass function (pmf) is a function that gives the probability that a **discrete random variable** is exactly equal to some value. The probability mass function is often the primary means of defining a discrete probability distribution.

Random variables that are discrete but largely unbounded (like number of web hits for a site each day) are often modeled with the so called Poisson distribution.

Example: probability mass function (pmf)

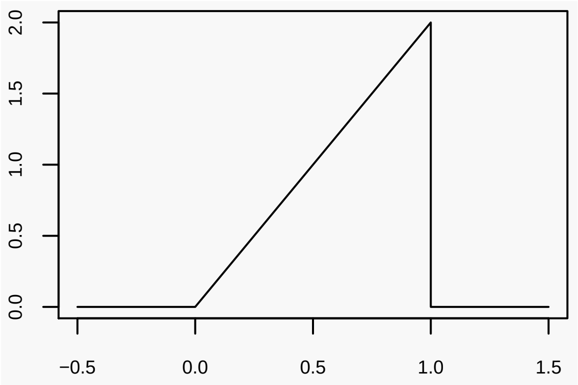
Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads. If the coin is biased and θ is the probability of a head expressed as a proportion (between 0 and 1):

$p(x) = \theta^x \cdot (1 - \theta)^{1-x}$

Example: probability density function (PDF)

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by:

$f(x) = 2x$ for $0 < x < 1$



Cumulative distribution function (CDF) and survival function

In addition to probability distribution functions, all random variables (discrete and continuous) have a cumulative distribution function (CDF, and p-prefixed functions in R). The CDF is a function giving the probability that the random variable X is less than or equal to x, for every value x, and models the accumulated probability up to that value. For continuous variables it's called density function:

$F(x) = P(X \leq x)$

The survival function (p-prefixed functions using `lower.tail = FALSE`) of a random variable X is defined as the probability that the random variable is greater than the value x:

$S(x) = P(X > x)$

Quantiles

The α^{th} quantile of a distribution with distribution function F is the point x_α so that:

$F(x_\alpha) = \alpha$

A percentile is simply a quantile with α expressed as a percent rather than a proportion. The (population) median is the 50^{th} percentile.

Miscellaneous

- In R, a prefix `p` returns probabilities, `d` returns the density, `q` returns the quantile and `r` returns generated random variables.
- By convention, we use an upper case X to denote a random, unrealized, version of a random variable and a lowercase x to denote a placeholder for a specific number that we plug into.

Afegitó

- Funcions amb prefix `d` a R → probabilitat **puntual** d'aquell valor x (substitueixen la x a la fórmula de la funció PDF/pmf).
- Funcions amb prefix `p` a R → probabilitat **acumulada** d'aquell valor x (substitueixen la x a la fórmula de la funció CDF).

3. Conditional probability

Let B be an event so that $P(B) > 0$. Then the conditional probability of an event A given that B has occurred is:

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

If two events A and B are unrelated in any way, or in other words independent, then the probability of both ocurring is:

$$P(A \cap B) = P(A)P(B)$$

And so, the conditional probability of an event A given that B has occurred if A and B are unrelated in any way is:

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

Finally, two events A and B are independent if $P(A|B) = P(A)$ and $P(B|A) = P(B)$. But don't confuse independent and disjoint or mutually exclusive. Disjoint and mutually exclusive mean the same thing, but independence is a very different concept!

Bayes' rule

Bayes' rule allows us to switch the conditioning event, provided a little bit of extra information:

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

Diagnostic tests

Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively. Let D and D^c be the event that the subject of the test has or does not have the disease respectively.

More or less easy to estimate:

- **Sensitivity**: probability that the test is positive given that the subject actually has the disease.
  $P(+|D)$
- **Specificity**: probability that the test is negative given that the subject does not have the disease.
  $P(-|D^c)$
- **Prevalence of the disease**: marginal probability of disease.
  $P(D)$

Difficult to estimate (so we'll use Bayes' rule here):

- **Positive predictive value**: probability that the subject has the disease given that the test is positive.
  $P(D|+)$
- **Negative predictive value**: probability that the subject does not have the disease given that the test is negative.
  $P(D^c|-)$

Diagnostic likelihood ratios

Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes.

Diagnostic likelihood ratios provide a way to summarize the evidence without appealing to an often unknowable prevalence. They summarize the evidence of disease given a positive or negative test:

- **Diagnostic likelihood ratio of a positive test**: $sensitivity/(1 - specificity)$.
  $DLR_+ = P(+|D)/P(+|D^c)$
- **Diagnostic likelihood ratio of a negative test**: $(1 - sensitivity)/specificity$.
  $DLR_- = P(-|D)/P(-|D^c)$

To interpret this results we must remember that if p is a probability, then $p/(1-p)$ is the odds. With this in mind, we calculate $P(D|+)/P(D^c|+)$ using Bayes' rule and we get this:

$$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+\ |D)}{P(+|\ D^c)} \times \frac{P(D)}{P(D^c)}$$

Or in other words:

post-test odds of disease $= DLR_+ \times$ pre-test odds of disease

So, DLR_+ is a multiplier saying "no matter what your pre-test odds of disease were (*often unknowable prevalence*), after a positive test your odds are now DLR_+ times those pre-test odds". Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

4. Variation

Remember the variance of X is defined as:

$Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$

The square root of the variance is called the standard deviation. The main benefit of working with standard deviations is that they have the same units as the data, whereas the variance has the units squared.

Example: variance from the result of a toss of a die

$E[X] = 3.5$

$E[X^2] = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.17$

$Var(X) = E[X^2] - E[X]^2 = 15.17 - 3.5^2 \approx 2.92$

Example: variance from the result of the toss of a potentially biased coin

$E[X] = 0 \cdot (1 - p) + 1 \cdot p = p$

$E[X^2] = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$

$Var(X) = p - p^2 = p(1 - p)$ *(This is a well known formula, so it's worth committing to memory.)*

Sample variance (n - 1 here!)

All the above is for the population variance. Sample variance is calculated this way instead:

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n - 1}$$

Why do we divide by $n - 1$ instead of n? To answer this we have to think in the terms of simulations. Remember that the sample variance is a random variable, thus it has a distribution and that distribution has an associated population mean. That mean is the population variance that we're trying to estimate if we divide by $n - 1$ rather than n. So the reason we use $n - 1$ rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance σ^2.

Thus, when talking about population variance, we divide by n and for a sample variance we use $n - 1$ instead.

This said, R always uses $n - 1$:

```r
dice <- c(1, 2, 3, 4, 5, 6)
sum((dice - mean(dice))^2)/6
## [1] 2.916667
```

```r
sum((dice - mean(dice))^2)/5
## [1] 3.5
```

```r
sd(dice)^2
## [1] 3.5
```

The standard error of the mean

We know that the expected value for the sample mean is the same as the population mean: $E[\bar{X}] = \mu$. How can we estimate the variability of the mean of a sample, when we only get to observe one realization?

The variance of the sample mean is $Var(\bar{X}) = \sigma^2/n$ where σ^2 is the variance of the population being sampled from.

This is very useful, since we don't have repeat sample means to get its variance directly using the data. We already know a good estimate of σ^2 via the sample variance. So, we can get a good estimate of the variability of the mean, even though we only get to observe one mean.

Often we take the square root of the variance of the mean to get the standard deviation of the mean. We call the standard deviation of a statistic its standard error (so in this case, $\sigma/\sqrt{n}$).

So, basically (and in R syntax):

- `var(x)` and `sd(x)` discuss how variable some x is in general.
- `var(x) / n` and `sd(x) / sqrt(n)` discuss the precision of our estimate of the mean of that x.

Basic properties for variance

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged:

$Var(X + a) = Var(X)$

If all values are scaled by a constant, the variance is scaled by the square of that constant:

$Var(aX) = a^2 \cdot Var(X)$

The variance of a sum of two random variables is given by:

$Var(aX + by) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$
$Var(aX - by) = a^2Var(X) + b^2Var(Y) - 2abCov(X, Y)$
$Cov(X, Y)$ is 0 when the two variables are independent from each other.

5. Some common distributions

Bernoulli (discrete): $X \sim Bernoulli(p)$

$P(X = x) = p^x(1 - p)^{1-x}$

The Bernoulli distribution arises as the result of a binary outcome, such as a coin flip. Thus, Bernoulli random variables take only the values 1 (typically called a "success") and 0 (a "failure") with probabilities of p and $1 - p$. Bernoulli random variables are commonly used for modeling any binary trait for a random sample.

$Mean = p$	$Variance = p(1 - p)$

The Bernoulli distribution is a special case of the binomial distribution where a single experiment/trial is conducted ($n = 1$).

Binomial (discrete): $X \sim Binomial(n, p)$

$P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$

The binomial random variables are obtained as the sum of iid Bernoulli trials. So if a Bernoulli trial is the result of a coin flip, a binomial random variable is the total number of heads (k successes).

Remember $\binom{n}{k}$ reads "n choose k" and is the number of ways of selecting k items out of n without replacement disregarding the order of the items and we calculate it as $\frac{n!}{k!(n-k)!}$. So $\binom{n}{k}$ is just the number of combinations $C(n, k)$. Thus, we are actually calculating the probability of having k successes times the number of combinations for getting those k successes.

$Mean = np$	$Variance = np(1 - p)$

```
# Probability of getting 7 or more girls out of 8 births?
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
## [1] 0.03515625
```

Poisson (discrete): $X \sim Poisson(\lambda)$

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space. For example, the number of meteors greater than 1 meter diameter that strike earth in a year, or the number of patients arriving in an emergency room between 10 and 11 pm, or the number of typing errors on a page, or the instances of mutation or recombination in a genetic sequence, or the distribution of errors produced in a sequencing process, or the probability of random sequence matches, or a process of radioactive decay. In a Poisson distribution, mean and variance have to be the same, which is something we can check provided we have enough data.

The Poisson distribution is an appropriate model if the following assumptions are true:

- The number of times an event occurs in an interval can take values $0, 1, 2, \ldots, \infty$.
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.

- The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
- Two events cannot occur at exactly the same instant.
- The probability of an event in a small interval is proportional to the length of the interval.

$Mean = \lambda$	$Variance = \lambda$

In a random process (such as mutation) there will be λ events per unit time interval. When dealing with rates (counts that occur over units of time), we use $X \sim Poisson(\lambda t)$ where $\lambda = E[X/t]$ and t is the total monitoring time.

Bernoulli, binomial and multinomial distributions can all be modeled by clever uses of the Poisson. For instance, when n is large and p is small, the Poisson is an accurate approximation to the binomial distribution (with $\lambda = np$).

Normal (continuous): $X \sim N(\mu, \sigma^2)$

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

$Mean = \mu$	$Variance = \sigma^2$

When $\mu = 0$ and $\sigma = 1$ the resulting distribution is called the standard normal distribution. Standard normal random variables are often labeled Z. We can transform normal random variables to be standard normals and vice versa:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

```r
# If BMI is a normal (mean 29 kg/mg2; sd 4.73), what is the population 95th percentile?
qnorm(.95, 29, 4.73)
## [1] 36.78016
```

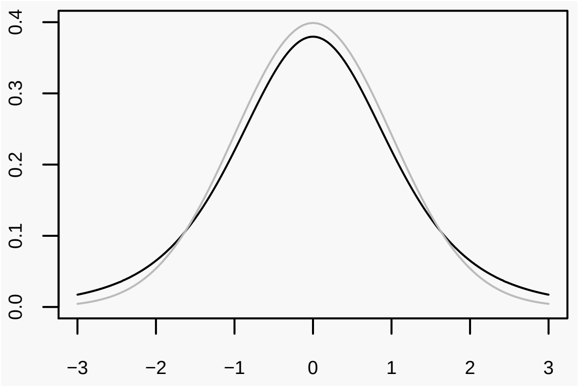
For the normal distribution, the location and scale parameters correspond to the mean and standard deviation, respectively. However, this is not true for most other distributions.

Many times, especially when dealing with physical phenomena (as opposed to humanly generated measurable data) the data will not be normally distributed. For non-normally distributed continuous data modeling we have two important families of distributions: the gamma family and the beta family. The gamma family consists of a few related distributions including the gamma distribution, the exponential distribution and the Chi-Square distribution (although the last two are special cases of the gamma distribution).

t-Student (continuous): $X \sim T(\nu)$

This distribution describes the sampling distribution of the sample mean when the true population variance is unknown, as is usually the case with sampling. In R we use `dt()` (and `t.test()` if we're calculating intervals).

$Mean = 0$ for $\nu > 1$	$Variance = \frac{\nu}{\nu-2}$ for $\nu > 2$



A t-distribution is a modification of the standard normal distribution to account for the variability of the standard deviation. So a standardized t-score is:

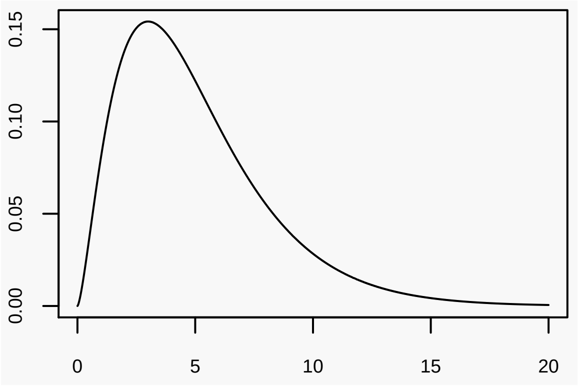
$t = \dfrac{X - \mu}{s}$

The t-distribution also applies to the sampling distribution of sample means as follows:

$t = \dfrac{\bar{X} - \mu}{s(\bar{X})} \qquad \text{with } s(\bar{X}) = \dfrac{s}{\sqrt{n}}$

If the data values $x_1, ..., x_n$ follow a normal distribution, then we call the distribution of the corresponding t scores a t-distribution with $n - 1$ degrees of freedom.

Chi-Square (continuous): $Q \sim X_k^2$



$Mean = k$	$Variance = 2k$

Being $k = n - 1$ the usual degrees of freedom. In R we use `dchisq()` (or `qchisq()` to produce Chi-Square test statistic values).

Gamma (continuous): $X \sim \Gamma(k, \theta)$

$P(X = x) =$ *...somewhat irrelevant...*

It provides a versatile model for working with continuous data that may not be normally distributed. Popular applications of the gamma distribution are to measurements of time until failure, concentrations of pollutants, etc. For the gamma distribution, h is the shape parameter and θ is the scale parameter.

$Mean = h\theta$	$Variance = h\theta^2$

The Chi-Square distribution seen before is actually a special case of the gamma distribution that always takes scale $\theta = 2$ and shape $h = k/2$, where k is the usual degrees of freedom.

$Mean = k$	$Variance = 2k$

The square of a standard normal random variable follows a Chi-Square distribution with one degree of freedom.

Exponential (continuous): $X \sim Exp(\lambda)$

The exponential distribution (`dexp()` in R) is famous for modeling survival times (as in the case with radioactive decay rates, or survival rates of bacteria or something of that sort) and is just a special case of the gamma distribution where the shape parameter k is 1. The exponential is often written in terms of a rate parameter λ where $\lambda = 1/\theta$:

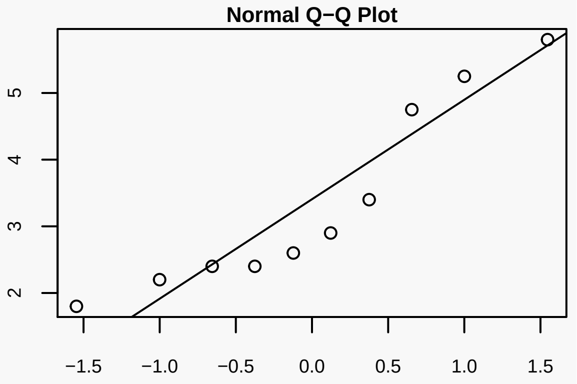
$Mean = \lambda^{-1}$	$Variance = \lambda^{-2}$

Example: Normal vs Gamma

Suppose we are measuring survival times of an enzyme in a solution and we get the following data in hours: 4.75, 3.4, 1.8, 2.9, 2.2, 2.4, 5.8, 2.6, 2.4, and 5.25. How could we decide on a probability model to model the probability of the enzyme surviving in solution?

Because we know we cannot always assume data is normally distributed, the first thing to do is to look at a plot of the data. There is actually a statistician's secret tool to check whether data are normally distributed. It is a plot called a Q-Q plot and what it does is line quantiles of the data against normal quantiles. If the line is a straight line, the data can be considered normally distributed and we can use the normal probability model.

```r
x <- c(4.75, 3.4, 1.8, 2.9, 2.2, 2.4, 5.8, 2.6, 2.4, 5.25)
qqnorm(x)
qqline(x)
```



It does not align nicely with the expected normal line. So we have to find a model other than the normal distribution to model the distribution of this data and gamma is a nice try as it's very flexible. To determine its shape and scale we know that k is the shape parameter and θ is the scale parameter:

```
mean(x) # k·theta
## [1] 3.35
```

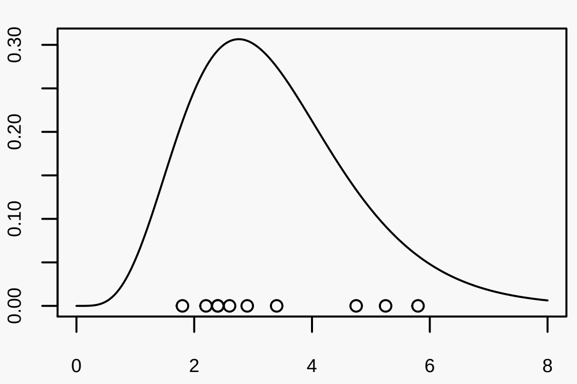
```
var(x)  # k·theta^2
## [1] 1.985556
```

So, basically:

```
k <- mean(x)^2 / var(x)
theta <- mean(x) / k
k; theta
## [1] 5.652071
## [1] 0.5927032
```

And we can finally plot the gamma distribution:

```
n <- length(x)
z <- seq(0, 8, length = 200)
plot(z, dgamma(z, shape = k, scale = theta), type = 'l')
points(x, rep(0, n))
```

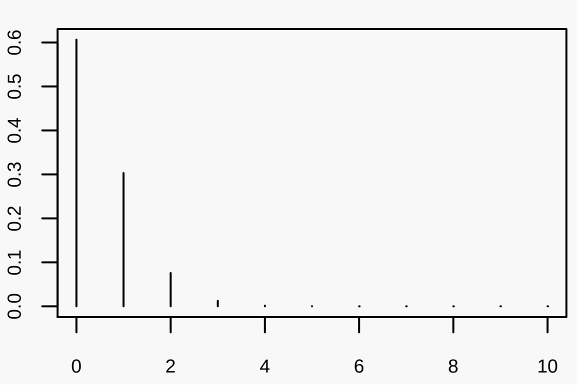


Afegitó

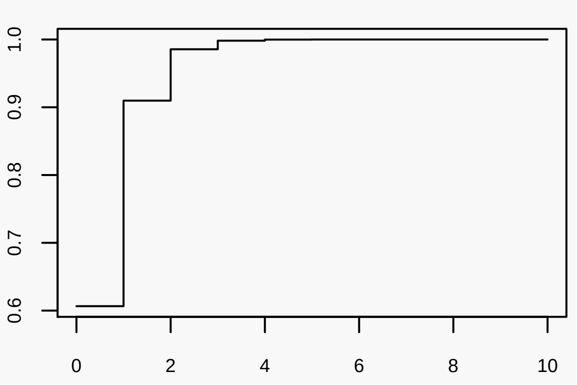
- Binomial → quan partim d'un percentatge d'èxit (p) dins la mostra.
- Poisson → quan és quelcom semblant a nombre d'arribades per unitat de temps (λ).
- Normal → quan partim d'una mitjana (μ).
- Binomial negativa → es comptabilitzen el número de successos 'negatius' que s'han de produir abans d'obtenir un número determinat de successos 'positius' k. A R, `dnbinom()` (i `dgeom()` per la distribució geomètrica, que és el cas concret de la binomial negativa quan $k = 1$).

A R, podem fer representacions bàsiques d'una funció utilitzant funcions `d` amb `type = 'h'` i funcions `p` amb `type = 's'` (per funcions contínues, `type = 'l'` en tots els casos).

```
# Probability
plot(0:10, dpois(0:10, 0.5), type = 'h')
```



```
# Cumulative Probability
plot(0:10, ppois(0:10, 0.5), type = 's')
```



6. Asymptotics

Asymptotics is the term for the behavior of statistics as the sample size limits to infinity (but it generally gives no assurances about finite sample performance). For example, for the mean the **Law of Large Numbers** says that if we go to the trouble of collecting an infinite amount of data, we estimate the population mean perfectly. An estimator is called consistent if it converges to what we want to estimate. Thus, the LLN says that the sample mean of iid sample is consistent for the population mean. The sample variance and the sample standard deviation of iid random variables are consistent as well.

In probability theory, the **Central Limit Theorem** (commonly known as the Law of Averages) establishes that, for the most commonly studied scenarios, when independent random variables are added, their sum tends toward a normal distribution even if the original variables themselves are not normally distributed. Thus, the distribution of averages of iid variables becomes that of a standard normal as the sample size increases:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{estimate - mean\ of\ estimate}{standard\ error\ of\ estimate} \sim N(0,1) \quad for\ large\ n$$

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

Confidence intervals

When we obtain a point estimate for a parameter we only have a single value. That value in and of itself says nothing about how accurate or precise the estimate is. Interval estimates provide an alternative method for estimating a parameter by providing a probable range of values the parameter is likely to take. Often it is good to use both a point estimate and an interval estimate for a parameter under study. The most common type of interval estimate is called a confidence interval:

point estimate ± standard error of point estimate × test statistic

The width of the interval depends on the standard error and the confidence level of the test statistic chosen. The test statistic comes from the appropriate sampling distribution of the parameter estimate (usually normal or t-distribution).

A confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (calculated from the observations), in principle different from sample to sample, that potentially includes the unobservable true parameter of interest. **How frequently the observed interval contains the true parameter if the experiment is repeated is called the confidence level**. In other words, if confidence intervals are constructed in separate experiments on the same population following the same process, the proportion of such intervals that contain the true value of the parameter will match the given confidence level.

Confidence intervals get wider as the coverage increases. Confidence intervals get narrower with less variability or larger sample sizes.

Taking the mean and adding and substracting the relevant normal quantile times the standard error yields a confidence interval for the mean. For example, 1.96 (or 2) times the standard error will yield a 95% interval.

How to construct a confidence interval using the CLT

The CLT let us consider this:

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

And we know that in a normal distribution, for 1.96 sds (or 2) we only leave 2.5% of values out. So we can say that the probability $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 2.5% each. Or equivalently, the probability that the interval with these limits contain μ is 95%:

$$\bar{X} \pm \frac{2\sigma}{\sqrt{n}}$$

The 95% refers to the fact that if one were to repeatedly get samples of size n, about 95% of the intervals obtained would contain μ. For a 90% interval we'd use 1.645 (and so on).

In binomial distributions, there's a quick fix to make an interval work better for small sample sizes: take your data and add two successes and two failures. So, for example, in our election example, we would form our interval with 58 votes out of 104 sampled (disregarding that the actual numbers were 56 and 100). This interval is called the Agresti/Coull interval and has much better coverage. In general, one should use the add two successes and failures method for binomial confidence intervals with smaller n.

Example: Poisson interval both using CLT and exact calculations

A nuclear pump failed 5 times out of 94.32 days. Give a 95% confidence interval for the failure rate per day.

Using asymptopia knowledge:

```r
x <- 5
t <- 94.32
lambda <- x/t
round(lambda + c(-1, 1) * qnorm(0.975) * sqrt(lambda/t), 3)
## [1] 0.007 0.099
```

Using exact calculations (Poisson and binomial cases have exact intervals that don't require the CLT):

```r
poisson.test(x, T = 94.32)$conf
## [1] 0.01721254 0.12371005
## attr(,"conf.level")
## [1] 0.95
```

Example: t-distribution confidence interval for two samples

We can create the confidence interval using:

$$\bar{Y} - \bar{X} \pm t_{df} \cdot S_p \cdot \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

Where t_{df} means $t_{n_x+n_y-2,1-\alpha/2}$ (that is a t quantile with $n_x + n_y - 2$ degrees of freedom) and the pooled variance estimator is:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

We want to compare the mean blood pressure between two groups in a randomized trial, those who received the treatment (Y) to those who received a placebo (X):

$$\begin{aligned} &\bar{Y} = 136.83\ mmHg \qquad s_y = 11.48\ mmHg \qquad (8\ people) \\ &\bar{X} = 126.66\ mmHg \qquad s_x = 15.68\ mmHg \qquad (21\ people) \end{aligned}$$

```r
# First we calculate the pooled variance estimate
ps2 <- ((21 - 1) * 15.68^2 + (8 - 1) * 11.48^2) / (21 + 8 - 2)
# Then we create the 95% confidence interval
136.83 - 126.66 + c(-1, 1) * qt(0.975, 27) * sqrt(ps2) * sqrt(1/21 + 1/8)
## [1] -2.367248 22.707248
```

We can see our interval contains 0, so we can't rule out 0 as the possibility for the population difference between the two groups. The same could be calculated in R:

```r
# x and y do exist and match n, mean and sd from the information above
t.test(y, x, var.equal = TRUE)$conf
## [1] -2.361288 22.705397
## attr(,"conf.level")
## [1] 0.95
```

In this example, we've used an unpaired analysis. However, paired differences t confidence intervals (`paired = FALSE` option in `t.test()`) are useful when:

- Pairs of observations are linked, such as when there is subject level matching or in a study with baseline and follow up measurements on all participants.
- There was randomization of a treatment between two independent groups.

How to calculate the standard error and coefficient for a confidence interval

For the standard error:

- If we do know the variance for the population σ^2, we simple calculate the standard error as $\frac{\sigma}{\sqrt{n}}$.
- If we do not know the variance for the population, then we use the variance for the sample S^2 and calculate the standard error as $\frac{S}{\sqrt{n-1}}$.

For the coefficient:

- If we do know the variance for the population, we use a z value from a standard Normal distribution.
- If we do not know the variance for the population, we use a t value from a Student's t-distribution with $n-1$ degrees of freedom.

How to calculate the required size of a sample

We can use this:

$margin\ of\ error = z \times standard\ error$

So, when dealing with a normal distribution it becomes:

$$me = z \cdot \frac{\sigma}{\sqrt{n}}$$

And when dealing with a binomial distribution it becomes:

$$me = z \cdot \sqrt{\frac{p(1-p)}{n}}$$

Being z the value from standard normal distribution to get the expected confidence level (1.96 for a 95%), σ an approximate expected value for the standard deviation and me (margin of error) the half of the desired interval's width. With this data we can solve the value for the required size of the sample (n).

As sample size increases, the margin of error decreases. As the variability in the population increases, the margin of error increases. As the confidence level increases, the margin of error increases.

Afegitó

- Per aproximar una binomial a una normal tirant del CLT, agafem $\mu = p$ i $Var = p(1-p)/n$. Per poder fer l'aproximació, cal que la mostra sigui prou gran, suficient com per fer $np > 5$ i $n(1-p) > 5$.

- A l'interval $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ hi haurà el 95% de les mitjanes de les mostres de mida n que s'agafin de la població (no confondre amb el 95% dels valors dins d'una mostra).

- Amb un interval de confiança el que diem és que el 95% dels intervals construïts amb $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ contenen la veritable mitjana poblacional μ.
- Exercici diapositiva 27 del document Estimación → Com no sabem la variança poblacional, cal aplicar t de Student pel coeficient i $\frac{S}{\sqrt{n-1}}$ per l'error estàndar. Per la t de Student, es busca $P(T \leq t) = 0.975$ amb 185 graus de llibertat (dóna 1.973).
- Exercici diapositiva 29 del document Estimación → El faig a partir d'una binomial, amb $n = 486$ i $r = 249$. Llavors, $p = r/n$ i error estàndar $\sqrt{\frac{p(1-p)}{n}}$. El coeficient ve d'una Normal (z).

7. Some common multivariable distributions

The joint distribution of two continuous random variables can be modeled using a joint probability density function (pdf). The joint pdf of two continuous random variables X and Y is a two dimensional area A and can be evaluated by integrating over this area with respect to each variable for given values of X and Y.

Calculating a marginal distribution of a continuous variable is similar to calculating a marginal discrete random variable distribution. In the case of the discrete random variable, this is done by summing over the other variable(s) whereas in the case of a continuous random variable, this is done by integrating over the other variable(s).

Distributions of more than one random variable are extensions of univariate distributions.

Multinomial (discrete)

Instead of just two possible outcomes (as in the case of the binomial), the multinomial models the case of multiple possible outcomes. For example, it models the probability of counts for rolling a k-sided die n times. If we count how many outcomes of each type occur, we have a set of k random variables X_1, X_2, …, X_k with values x_1, x_2, …, x_k (and the sum of all x_i is n).

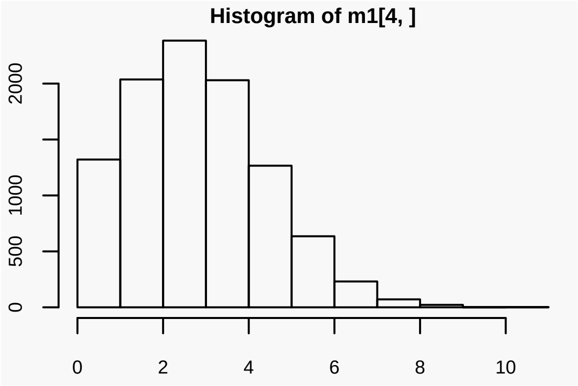
$P(x_1, x_2, ..., x_k) = \binom{n}{x_1 x_2 ... x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$

Remember the value for the multinomial coefficient is:
$\binom{n}{x_1 x_2 ... x_k} = \frac{n!}{x_1! x_2! ... x_k!}$

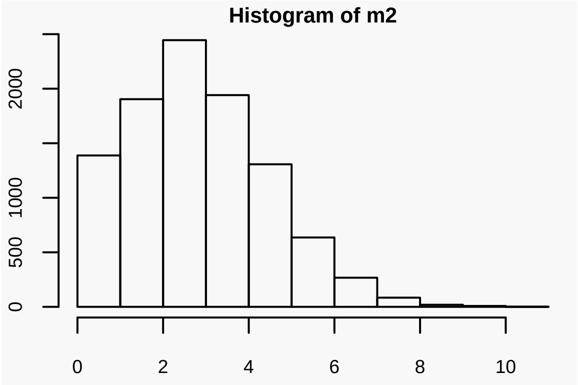
```
# Simulation of throwing 20 dice at once 10 times
rmultinom(n = 10, size = 20, prob = rep(1/6, 6))
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    4    3    2    2    4    3    2    3    2     7
## [2,]    3    5    4    1    2    2    4    5    6     5
## [3,]    2    6    6    5    2    0    2    2    2     2
## [4,]    6    3    1    6    3    4    5    3    3     1
## [5,]    3    1    5    2    2    7    2    3    4     3
## [6,]    2    2    2    4    7    4    5    4    3     2
```

If we are only interested in the number of outcomes that result for number 4, then we simply lump all the other results into one category called "other". Now we have reduced this to a situation we have two outcomes. This should ring a bell of familiarity, as it has now become a case of Bernoulli trials:

```
# Using multinomial distribution and taking marginal for 4
m1 <- rmultinom(n = 10000, size = 20, prob = rep(1/6, 6))
hist(m1[4,], breaks = 12)
```



```
# Using binomial distribution
m2 <- rbinom(n = 10000, size = 20, prob = 1/6)
hist(m2, breaks = 12)
```



Multivariate normal (continuous)

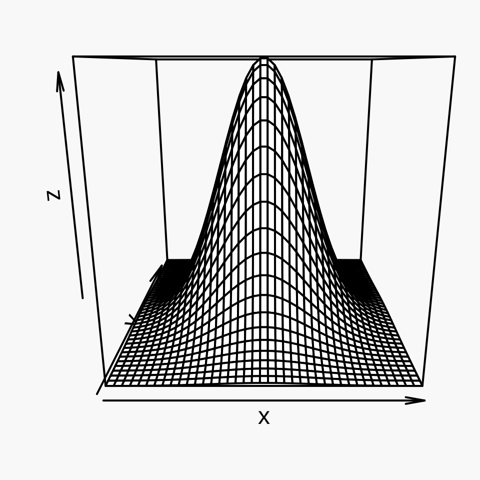
Most of inferential multivariate statistics utilize the multivariate normal. To avoid dealing with advanced techniques of calculus and matrix algebra, we will consider the details of only the bivariate normal model, which models the joint distribution of two independent normally distributed random variables.

$$P(X = x, Y = y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Considering both X and Y are standard normal random variables (mean $= 0$, sd $= 1$), this becomes:

$$P(X = x, Y = y) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{y^2}{2}} = \frac{1}{2\pi} \cdot e^{-\frac{(x^2+y^2)}{2}}$$

```
# Perspective for a bivariate standard normal distribution
len <- 40
x <- seq(-3, 3, length = len)
y <- x
z <- x %*% t(y)
for (i in 1:len) { for(j in 1:len) { z[i, j] <- (1/2*pi) * exp(-(x[i]^2+y[j]^2)/2) }}
persp(x, y, z)
```

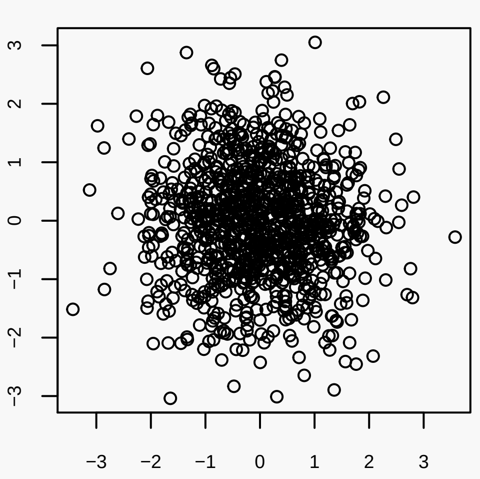


In R, we can use `rmvnorm()` function (`mvtnorm` package) to generate values for X (first column) and Y (second column) from a standard normal simulation:

```
# Generating 1000 values from a bivariate standard normal distribution
data <- rmvnorm(1000, mean = c(0, 0))
head(data)
##            [,1]       [,2]
## [1,] 1.1594888  0.4261927
## [2,] 2.4904938  1.3922035
## [3,] 0.2534997 -0.5562399
## [4,] 0.2466322 -1.8866914
## [5,] 0.1724956  0.3723070
## [6,] 0.6767905  0.7064631
```

And now we can use a scatter plot to check if the joint density is shaped as we would expect data from a bivariate normal simulation:

```
plot(data[,1], data[,2])
```



Finally, if we want to look at marginal distributions of X and Y, we just need to consider the first or second column, respectively.

8. Foundations of Statistical Inference

Sampling distributions are a special class of probability distributions where the shape of the curve changes based on the sample size n. The criteria "degrees of freedom" (based in part on sample size) is part of the defining parameter for plotting a sampling distribution. **Sampling distributions are distributions of statistics rather than distributions of individual data values.**

Every statistic has it's own sampling distribution. Here we consider the sampling distribution for the mean (the t-distribution) and the sampling distributions of statistics that are based on variance (the Chi Square, and the F distributions).

Mean with t-distribution

The t-distribution is specially well suited for small samples. As the degrees of freedom (sample size) increases, the distribution in the limiting case (n approaching infinity) becomes normal. In statistics this is referred to as an asymptotic approximation. Indeed when sample size is roughly 30 or so we often use the normal distribution instead of the t-distribution because of this close approximation.

Variance with Chi-Square

The Chi-Square distribution indirectly models the sample variance. The ratio of the sample variance to the true population variance is modeled as a Chi-Square according to the following:

$$\frac{ks^2}{\sigma^2} \sim \chi^2_k \quad \text{where } k = n - 1 \ (degrees\ of\ freedom)$$

Assuming normality, we can create a $100(1-\alpha)\%$ confidence interval for the variance σ^2 this way:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right)$$

Variance with F distribution

The F distribution models the distribution of the ratio of two independent random variables U and V, each having a Chi-Square distribution (with m and n degrees of freedom respectively). In R we use `df()` (and `qf()` for quantiles, i.e. in ANOVA).

$$W = \frac{U/m}{V/n} = \frac{nU}{mV}$$

The F distribution is an incredibly useful distribution and the basis for many statistical inference tests, as it creates a testable "signal to noise" ratio. Suppose the variability of background noise is modeled with random variable U and variability due to the experimental effects is modeled with random variable V. Computing the ratio of variability and modeling it with an F distribution provides a criterion to determine if the experimental procedure achieves a significant effect over the background noise level. F ratios serve as the basis for ANOVA (analysis of variance), which in turn is the basis for the branch of statistics involving the design of experiments.

Degrees of freedom

The term "degrees of freedom" is a description of the number of observations that are free to vary after the sample statistics have been calculated.

This is a brilliant explanation for this concept from stats.stackexchange.com talking about the variance:

> If you have the whole population at your disposal then its variance (**population variance**) is computed with the denominator n. Likewise, if you have only sample and want to compute this **sample's variance**, you use denominator n (n of the sample, in this case). In both cases, note, you don't *estimate* anything: the mean that you measured is the true mean and the variance you computed from that mean is the true variance.
>
> Now, you have only sample and want to *infer* about the unknown mean and variance in the population. In other words, you want *estimates*. You take your sample mean for the estimate of population mean (because your sample is representative), OK. To obtain estimate of population variance, you have to pretend that that mean is really population mean and therefore it is not dependent on your sample anymore since when you computed it. To "show" that you now take it as fixed you reserve one (any) observation from your sample to "support" the mean's value: whatever your sample might have happened, one reserved observation could always bring the mean to the value that you've got and which believe is insensitive to sampling contingencies. One reserved observation is "-1" and so you have $n-1$ in computing variance estimate.
>
> Imagine that you somehow know the true population mean, but want to estimate variance from the sample. Then you will substitute that true mean into the formula for variance and apply denominator n: no "-1" is needed here since you know the true mean, you didn't estimate it from this same sample.

Parameter estimation for points

Given our sample data we want to fit a model to the data. Often we would like to fit a standard probability model. In order to do this, we need to estimate the best fitting parameters for the particular model we have in mind. Parameter estimates take two forms: point estimates and interval estimates. Point estimates have their merit in being very useful in defining the model. Interval estimates have merit in quantifying how precise a parameter estimate is.

In any given sample the point estimate differs from the true underlying population parameter. If in many repeated (conceptual) samples this difference averages out to zero, we say that the point estimate is unbiased.

Some usual point estimates: $\bar{X}$ (dividing by n) for μ, s^2 (dividing by $n-1$) for σ^2 and s for σ in normal populations and p (successes / trials) for p in a binomial model.

But what if we want to estimate more complicated parameters like the parameters of a gamma distribution? There exist no simple calculations of point estimates of these population parameters. Therefore in many cases we need a more sophisticated method to make point estimates. Often the maximum likelihood method (MLE) works best.

Maximum Likelihood Estimation (MLE)

Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum. The maximum likelihood estimate for a parameter μ is denoted $\hat{\mu}$ and is calculated using some algebra and calculus (a maximum value occurs when the first derivative of a function is set to zero).

A maximum likelihood estimator is a value of the parameter a such that the likelihood function (the function of the parameters given the data) is a maximum.

For a Bernoulli or binomial distribution:

$\hat{p} = \dfrac{\sum x_i}{n}$ so $\frac{number\ of\ successes}{number\ of\ trials}$

For a normal distribution:

$$\hat{\mu} = \frac{\sum x_i}{n} \qquad \hat{\sigma} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Note that in this case, the maximum likelihood standard deviation is the sample standard deviation, which is a biased estimator for the population standard deviation (using $n - 1$ would do for an unbiased version).

For a Poisson distribution:

$$\hat{\lambda} = \frac{\sum x_i}{n}$$

So say we've got a data set of $y = -1, 3, 7$. The most consistent normal distribution from which this data could have come has a mean of 3 and a variance of 16. It could have been sampled from some other normal distributions, but one with a mean of 3 and variance of 16 is the most consistent with the data in the following sense: the probability of getting the particular y values we observed is greater with this choice of mean and variance than it is with any other choice.

9. Bootstrapping

Bootstrapping is a technique to find standard errors (measures of variability in the data) or confidence intervals in complicated situations where analytical computation is impossible (data do not fit a nice distribution pattern and also the law of averages does not work well for the data at hand). Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution.

Bootstrap techniques can be parametric or nonparametric. Parametric ones assume that the data are generated from a standard parametric probability model (normal, Poisson, etc). Because of their versatility, nonparametric bootstrap techniques are the more popular type of bootstrap applications.

The bootstrap estimates the standard error of a metric (such as a parameter estimate, mean, median, etc) by repeatedly drawing bootstrap samples from the original data. The samples are drawn with replacement and the sample size of each sample is the same as the original sample size. For each bootstrap sample, the metric is measured. Typically about 1000 bootstrap samples are taken. Then, the standard error of the metric under study is measured using the observed variation of the bootstrap samples. In R, there are bootstrapping functions in the `boot` package.