Further topics and Matlab for Probability and Statistics 2015

Johan Jonasson [*†‡]

December 2014

1 Simulation and some new distributions

In Matlab's Statistics Toolbox, there are ready-made functions for generating random numbers of most probability distributions that we have encountered. Therefore the first few of the following exercises may feel a bit superfluous. However, it is of value, also from a theoretical point of view, to know how to transform random numbers of one distribution to some other, desired, distribution.

Assume that F_1 and F_2 are two distributions functions. For simplicity, assume to begin with that F_1 and F_2 are both strictly increasing, which in particular entails that they are invertible.

Proposition 1.1 *Assume that X is a random variable with distribution function F_1. Let $Y = F_2^{-1}(F_1(X))$. Then Y has distribution function F_2.*

The truth of the proposition follows from

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}\left(F_1(X) \leq F_2(y)\right) = \mathbb{P}\left(X \leq F_1^{-1}(F_2(y))\right) \\ &= F_1(F_1^{-1}(F_2(y))) = F_2(y).\end{aligned}$$

A special case of Proposition 1.1 is that if $X \sim \text{unif}[0,1]$, then $F^{-1}(X)$ has distribution function F. This is very interesting, since computer generation of

[*]Chalmers University of Technology
[†]Göteborg University
[‡]jonasson@chalmers.se

random numbers usually starts from the uniform distribution. In Matlab, one generates a unif$[0,1]$ distributed random variable by `unifrnd(0,1)`. **Now try** to generate a vector `x` of 1000 such random numbers and visualize the result by plotting a histogram and the *empirical distribution function*[1]

Now try to simulate the exponential distribution and the normal distribution for some different parameters. The normal distribution function is nor possible to invert explicitly, but matlab has the function `norminv` for this. **The also find** the functions that directly generate data according to these distributions, and compare.

Note. Matlab parametrizes the exponential distribution with the expectation, i.e. $1/\lambda$.

When a distribution function of interest is not invertible, which is the case e.g. for all discrete distributions, then one can instead use the generalized inverse

$$\overline{F}(y) = \int \{x : F(x) \geq y\}.$$

Since F is right continuous, $F(\overline{F}(y)) = y$ for all y, the above proposition holds with F^{-1} replaced with $\overline{F}$.

Now, the distribution function of a discrete random variable has a staircas shape, so its generalized inverse has an awkward expression. Luckily, Matlab has functions for all the common discrete distributions. **Your job** is now to find these for the geometric, the Poisson and the binomial distributions. **Find also** the functions for directly generate data according to these and compare.

It is now time to learn a few new distributions.

Definition 1.2 *A random variable X is said to be gamma distributed with parameters $\alpha > 0$ and $\lambda > 0$, written $X \sim \Gamma(\alpha, \lambda)$, if it has the density*

$$f(x) = \frac{1}{C}\lambda^{\alpha} x^{\alpha-1} e^{-\alpha x},\ x \geq$$

where C is the normalizing constant $\Gamma(\alpha) := \int_0^{\infty} e^{-t} t^{\alpha-1} dt$.

If α is an integer, it is fairly easy to compute that $\Gamma(\alpha) = (\alpha - 1)!$. Taking $\alpha = 1$ gives exactly the density for an exponential random variable, i.e.

$$X \sim \Gamma(1, \lambda) \Leftrightarrow X \sim \exp(\lambda).$$

[1] If $x_1, x_2, \ldots, x_n$ is a set of numbers, i.e. a sample of some random variable, the empirical distribution function of the x_k:s is given by the distribution function of the random variable Z such that $\mathbb{P}(Z = x_k) = 1/n$ for all $k = 1, \ldots, n$

In fact, if $X_1, \mathcal{X}_2, \ldots, X_n$ are independent and $\exp(\lambda)$ distributed, then

$$\sum_{k=1}^{n} X_k \sim \Gamma(n, \lambda).$$

Prove this, using induction. An immediate consequence of this is that the time up to the n'th point in a Poisson process of intensity λ, is $\Gamma(n, \lambda)$ distributed. We also get the formulas

$$X \sim \Gamma(\alpha, \lambda) \Rightarrow \mathbb{E}[X] = \frac{\alpha}{\lambda}, \, \mathbb{V}\mathrm{ar}[X] = \frac{\alpha}{\lambda^2}$$

for integer α. The formulas are true for general α, which at this point is not hard to believe. (We will not prova this.) Not also that according to the Central Limit Theorem, $\Gamma(n, \lambda)$ is very close to $\mathrm{N}(n\lambda, (\sqrt{n}/\lambda)^2)$ for large integer n. **Now generate** vectors of $\Gamma(4, 1)$ distributed random numbers in three different ways: (i) transform as above, using `gaminv`, (ii) sum four independent exponential random numbers and (iii) using the function `gamrnd`. **Try also** to approximate the probability that the third point in a Poisson process of intensity 2 arrives after time 2.4.

In the light of the CLT, it is reasonable to assume a normal distribution for many random and naturally appearing quantities. In some cases, however, such an assumption is clearly unreasonable. In particular, this is the case for many, but far from all, situation where the quantity in question is obviously nonnegative. Since any normal distribution has its support on the entire real line, such a quantity cannot be *exactly* normally distributed. Sometimes the error we get from this is negligible, sometimes not. An example of the former kind is the following. Consider the height that a son of a mother who is 170 cm tall, will get as an adult. Say that the son's expected height is 184 cm with a standard deviation of 6 cm. The son's height is obviously positive, so it cannot be exactly normal. However, zero is more than 28 standard deviations below the expectation, so the error will be of order $e^{-28^2/2}$, a vanishingly small number.

Consider instead the following. A patient takes 75 mg of a drug and two hours later one measures the patient's blood concentration of the active substance. It is well known that the metabolism of drugs may vary very much between different persons, so we may very well have a situation where the expectation is, say, 120 (ng/ml) and the standard deviation is 80. Here the assumption of a normal distribution will then cause an unacceptable error. In situations like these, one usually can instead assume that the *logarithm* of X is normal, i.e. one can write $X = e^Y$ where Y is normal.

Definition 1.3 *Let $Y \sim N(\mu, \sigma^2)$ and $X = e^Y$. Then X is said to have a lognormal distribution with parameters μ and σ^2, $X \sim logN(\mu, \sigma^2)$.*

Observe that is *not* true that $X \sim \text{logN}(\mu, \sigma^2) \Rightarrow \mathbb{E}[X] = e^\mu$; in fact $E[X] > e^\mu$ unless $\sigma = 0$. However if $\sigma << \mu$, then the difference is very small. Shouldn't one always assume a lognormal rather than a normal distribution for positive quantities, like e.g. the son's height above? Sure, but if $\sigma << \mu$, the difference gets so small that it is not worth the trouble![2]

Exercise: Find the Matlab function the generates lognormal random numbers and check that you get the same result as when you generate normal Y:s and then take e^Y. If $X \sim \text{logN}(0, 1)$, what does $\mathbb{E}[X]$ seem to be? More than e^0, doesn't it?

Another distribution that arises from the normal distribution is the χ^2 distribution.

Definition 1.4 *Let $Z_1, Z_2, \ldots, Z_n$ be independent and standard normal and let*

$$X = \sum_{k=1}^{n} Z_k^2.$$

Then X is said to be χ^2-distributed with n degrees of freedom and one writes $X \sim \chi_n^2$.

It follows immediately that if $X_1, X_2, \ldots, X_n$ are independent and $\text{N}(\mu, \sigma^2)$ distributed, then

$$\frac{\sum_{k=1}^{n}(X_k - \mu)^2}{\sigma^2} \sim \chi_n^2.$$

If Z is standard normal, then $\mathbb{E}[Z^2] = 1$. From this it follows that $X \sim \chi_n^2 \Rightarrow \mathbb{E}[X] = n$. One can also (can you?) show that $\mathbb{V}\text{ar}[X] = 2n$.

In fact, perhaps surprisingly, the χ^2 distribution is a special case of the gamma distribution:

$$X \sim \chi_n^2 \Rightarrow X \sim \Gamma(n/2, 1/2).$$

In particular, if Z_1 and Z_2 are independent and standard normal, $Z_1^2 + Z_2^2 \sim \exp(1/2)$. **Prove this.** Then **use this result** to simulate standard normal random

[2]This implies that sometimes X and $\log X$ can both be assumed to be normal. Isn't this a contradiction? Yes, it is not possible for them both to be *exactly* normal, but if $\sigma << \mu$, then they can both be very close to normal.

numbers, using the exponential distribution. Hint: Regard Z_1 and Z_2 as the coordinates of a random vector (Z_1, Z_2) and consider the polar coordinates R and θ. What you just showed is that $R^2 \sim \exp(1/2)$. One can also show that R and θ are independent and $\theta \sim \text{unif}[0, 2\pi]$. Now you can simulate R and θ and express Z_1 in terms of these.

Another very important fact is

Proposition 1.5 *If $X_1, \ldots, X_n$ are independent and N(μ, σ^2) distributed and s^2 is as usual, defined as*

$$s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_i - \overline{X})^2,$$

then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We will not prove this. **Simulate** to convince yourself. Take e.g. $n = 10$, repeat 1000 times and compare with 1000 observations generated by Matlab's function `chi2rnd`. (To compute the sample variance for a data vector is simple: just use `var`.)

2 Failure rates and some more distributions

Let f be the pdf of a continuous random variable $X \geq 0$. Recall that one interpretation of the pdf is that if $h > 0$ is a small number, then $\mathbb{P}(X \in (x, x+h)) \approx hf(x)$. Let $G(x) = 1 - F(x) = \mathbb{P}(X > x)$. We then get

$$\mathbb{P}(X \in (x, x+h) | X > x) = \frac{\mathbb{P}(X \in (x, x+h))}{G(x)} \approx h \frac{f(x)}{G(x)}.$$

If X is interpreted as the life length of something, then this can be loosely spelled out at as the "probability to fail in the next moment given survival up to now". It is therefore logical to define the ration on the right hand side as the *failure rate* or *death rate* of X.

Definition 2.1 *The failure rate of X is given by*

$$r(x) = \frac{f(x)}{G(x)}, x \geq 0.$$

Observe that if one knows r, then one can compute G (and hence the distribution of X); since $r(x) = -G'(x)/G(x) = -(d/dx)\log G(x)$, we get

$$G(x) = \exp\left(-\int_0^x r(t)dt\right).$$

Suppose that X has a constant failure rate, λ. Then we get $G(x) = e^{-\lambda x}$, i.e. X is exponential with parameter λ. This should come as no surprise, since we know the exponential distribution as the life length of something that does not age with time. This is often a reasonable assumption when dealing e.g. with electrical components, but unreasonable for many mechanical components and clearly unreasonable for the life length of animals or humans. It also turns out to be unrealistic in most situations when one is dealing with data that are extreme values of some kind, e.g. the highest sea level of the month or the strongest wind of the decade. We will now encounter two sets of data where the exponential distribution is not a realistic model.

Download the file `atlantic.dat`. It contains real data in the form of 582 so called significant wave heights, measure in the Atlantic. The significant wave height is defined as the average of the highest third of the waves during the period allotted for that measurement. Data are registered 14 times per month. You can download the file e.g. by typing `y=load('atlantic.dat')` and will then get the observations in the vector `y`. Use the data to estimate the failure rate, using the following two lines

```
for i=1:100, G(i)=length(y(y>0.12*i))/582;, end

for i=1:99, r(i)=(G(i)-G(i+1))/(0.12*G(i));, end
```

Here we have partitioned the interval $[0, 12]$ into one hundred subintervals and first estimated $G(x) = 1 - F(x)$ and then $f(x)/G(x)$. Plot this and see how it looks. Does the failure rate seem to be increasing or decreasing or neither?

Definition 2.2 *The random variable $X \geq 0$ is said to be Weibull distributed, with shape parameter k and scale parameter λ if its distribution function F is given by*

$$F(x) = 1 - e^{(\lambda x)^k}, x \geq 0.$$

By differentiating, we find that the pdf is given by

$$f(x) = k\lambda(\lambda x)^{k-1}\exp(-(\lambda x)^k).$$

Now compare you estimated failure rates with the failure rates for the Weibull distribution for the parameters that fit best with the data. You can find these parameters by using the function `wblfit`. Observe that Matlab has a parametrization of the Weibull distribution that differs from ours; our (k, λ) corresponds to $(1/\lambda, k)$ in Matlab.

What you get shouldn't look too bad, apart from the fact that the estimated failure rate is so erratic, that it is hard to judge how good the fit really is. This is typical for estimated failure rates, so, to be true, comparing failure rates is not a very good way to compare distributions.

A more efficient way is to compare the empirical distribution functions of the data with the empirical distribution function for a set of data generated according to distribution you want to compare with (in this case the Weibull distribution with the best parameters). **Do this** for the wave data and plot in the figure with `cdfplot`. A much clearer picture, right?

Even better is of course to compare data with the exact distribution function of the distribution you compare with. This can be done with a *probability plot*. A probability plot is a plot of the data in a coordinate system where the axes have been re-scaled, so that if data fits well with the distribution you compare with, then data will be close to a straight line. Such ready-made probability plot functions are installed in Matlab for a few standard distributions, the normal distribution and the Weibull distribution among them. The function for a Weibull plot is `wblplot`. **Do a Weibull plot.** The fit should look reasonable, but not perfect. In particular you should see that the fit is rather poor in the tails.

Definition 2.3 *The random variable X is said to be Gumbel distributed with scale parameter a and location parameter b, $X \sim$ Gumb(a, b), if*

$$F(x) = \exp(-e^{-(x-b)/a}), x \in \mathbb{R}.$$

Compare the wave data with a Gumbel distribution in the same way as for Weibull; by comparing with simulated data from Gumbel and by making a Gumbel probability plot. Matlab does not have ready-made functions for the Gumbel distribution, but you can download the files `gumbfit`, `gumbcdf` and `gumbplot` from the course url. Hopefully you will find that wave data fit better with a Gumbel distribution than with Weibull.[3]

[3]Here again we have data that are necessarily positive, but the Gumbel distribution has its support in the entire real line. Again, this can be OK if the standard deviation is much smaller than the expectation.

An important property of the Gumbel distribution is that it is *max stable*, which is to say that if X_1 and X_2 are independent and Gumbel with the same scale parameter, then $\max(X_1, X_2)$ is also Gumbel. **Prove this.** (What is the location parameter of the maximum?) This is one reason why the Gumbel distribution tends to fit well with some extreme value data.

Let us now look at failure rates for a different collection of data, concerning life length of humans. Download data from the file `norway.mat` by typing `load 'norway.mat'`. You will get the data in the form of the matrix `norway`. In the second column of this matrix, one finds estimated life lengths of Norwegian females and in the third column one finds the corresponding figures for Norwegian males. These data are to be interpreted that if the number in in the n'th entry is x, then, out of 100000 hypothetical (fe)males born in the year 2000, x of them survive the their $n-1$'th birthday. Use these data to estimate the failure rates for Norwegian women/men. Plot them both in the same figure. Do you see a difference? (Since the data concerns hypothetical persons, the estimated failure rates will be nowhere near as erratic as in the example above.)

Now check if data fits with a normal, Weibull or Gumbel distribution. No good, right? In fact, data fits much better with a model that is actually used by life insurance companies, namely with a failure rate given by

$$r(t) = a + be^{-ct}.$$

In the present case, one has $a \approx 9 \cdot 10^{-4}$, $c \approx 10.3$ and $b = 3.3 \cdot 10^{-5}$ for females and $b = 4.4 \cdot 10^{-5}$ for males. Compute, for this model with these a, b and c, the conditional probability that a Norwegian man born in 2000, becomes at least 80 years, given that he becomes at least 30 years.

3 The Poisson process

The Poisson process consists of a set of occurrences in time, such that the time gaps between consecutive occurrences (including the time from the start to the first occurrence) are independent and exponentially distributed with a common parameter λ. The parameter λ is then called the *intensity* of the Poisson process. It is straightforward to simulate a Poisson process; just let $T_1, T_2, \ldots$ be independent $\exp(\lambda)$ random variables and let S_n, the time for the n'th occurence, be $S_n = T_1 + \ldots + T_n$. Write $X(s,t)$ for the number of occurrences in the time interval (s,t) and write $X(t)$ for $X(0,t)$. We have seen (or will see) that $X(s,t)$ is Poisson

distributed with parameter $\lambda(t-s)$. **Try this out** by simulating as described above and check how many occurrences you get in $[2,5]$. Use $\lambda = 2$. Repeat for 1000 times and compare to what you get if you simulate directly according to the expected Poisson distribution.

Two important properties of the Poisson process are the *superposition* property and the *thinning* property. The superposition property says that if $X(t)$ and $Y(t)$ are two independent Poisson processes with intensities λ_x and λ_y respectively, then $X(t)+Y(t)$ is a Poisson process of intensity $\lambda_x+\lambda_y$. We call $X(t)+Y(t)$ for the superimposed Poisson process and it is the process that takes the union of occurrences in $X(t)$ and $Y(t)$ as occurrences. **Try this out** by simulating a superimposed Poisson process, register the times between occurrences and compare with the expected exponential distribution.

Observe that the superposition property is equivalent to the that if Z_1 and Z_2 are independent and $\exp(\lambda_x)$ and $\exp(\lambda_y)$ respectively, then $\min(Z_1, Z_2)$ is exponential with parameter $\lambda_x+\lambda_y$.

The thinning property says that if $X(t)$ is a Poisson process with intensity λ and $Y(t)$ is the process that takes for occurrences each occurrence of $X(t)$ but only with probability p independently. Then $Y(t)$ is a Poisson process of intensity λp. **Try this out too** as for the superposition property.

The Poisson process for all types of processes where it seems reasonable to assume that, at all times, neither time since the last occurrence nor the time in itself, affects the distribution of the time to the next occurrence. Examples of processes of occurrences of this kind are road accidents, storms, earth quakes, goals in a football game, groups of people you meet on your evening walk, radioactive decay, etc. Let us take a look at an example of real data. **Download** the file `coal.dat`. The file contains information about fatal accidents in British coal mines from 1851 to 1918, in the form of a matrix with six columns, where the columns contain the following information about the accidents.

1. Day of month.
2. Month.
3. Year.
4. Number of the day of that year.
5. Number of day since the last fatal accident.

6. Number of fatal accidents.

Have the fatal accidents arrived as a Poisson process? Test this by comparing the times between consecutive accidents with a sample of the exponential distribution. You should expect to get a fairly good fit, but a closer look should also reveal that there seem to be "too many" very short and very long times between occurrences. One can then suspect that what we have really observed is a *time inhomogeneous* Poisson process, i.e. a Poisson process where the intensity changes over time. (In such a process, at a given time, the distribution of the time to next occurrence is not affected by the time since the last occurrence, but it is affected by the time itself.) We will not be going into more detail about Poisson processes in this more generals sense, more than just stating that the time homogeneous case (i.e. what we have studied so far in this course) is a convenient special case. Most of the processes that one models as a Poisson process will be time inhomogeneous if studied over a long enough time span, but it is often reasonable to assume time homogeneity over shorter time scales.

Looking more closely at the coal mine data, one can see a change in intensity around accident no. 127. **Study data** before and after this separately to see if these fit better with a homogeneous Poisson process and if there indeed seems to be an essential difference in intensity. (Around 1880, there was a legislation effort to improve safety in British coal mines. Does it seem that this had an effect?)

4 Tests for normally distributed data

A "standard situation" is to make confidence intervals and tests under the assumption that the observations are distributed according to a normal distribution. We have seen this be done for one sample as well as two samples. In the one sample case, one assumes that one has access to a sample $X_1, \ldots, X_n$ with $X_k \sim \mathrm{N}(\mu, \sigma^2)$, where μ is unknown and σ may be known or unknown. Let us assume here the most common case that σ is unknown. We want to make a confidence interval for μ os test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_A : \mu \neq \mu_0$ or $H_A : \mu > \mu_0$ or $H_A : \mu < \mu_0$. Recall the correspondence between confidence intervals and tests: H_0 is rejected at significance level α if and only if the corresponding confidence interval for μ at confidence level $1 - \alpha$ does not contain μ_0. Here the confidence interval

$$\mu = \overline{X} \pm F^{-1}_{t_{n-1}}(1 - \alpha/2)\frac{s}{\sqrt{n}}$$

corresponds to the alternative hypothesis $H_A : \mu \neq \mu_0$, whereas $H_A : \mu > \mu_0$ corresponds to the lower bounded confidence interval

$$\mu \geq \overline{X} - F^{-1}_{t_{n-1}}(1-\alpha)\frac{s}{\sqrt{n}}.$$

Here too, we will work with real data. **Download** the file `birth.dat`: in the Matlab window, click in "Import data", choose "birth" and then download. Data consists of a 747×26-matrix containing lots information about newborn babies of mothers that at four maternity clinics in Malmö during the years 1991-1992. You will find information about the content in the different columns of `birth` by reading the file birth.txt. (Note that I have replaced the original "." for "missing data" with "99" in order to make Matlab accept the data.)

We will be mostly interested in making inference about the weight of the children at birth. Create a vector `fv=birth(:,3)` containing the weights at birth. We want to assume that these are normal. **Make a normal distribution plot.** Does it look OK? It should look pretty good, but not perfect. In particular it should be evident from the plot that we have "too many" very low weights than expected from the normal distribution. In any case, it does not look too bad, so let us assume that data are indeed normal.

As a rule of the thumb, one usually says that the weight of of a newborn Swedish infant has an expectation of 3500 g, with a standard deviation of around 500 g. **Make a test** on 5% significance level of $H_0 : \mu = 3500$ against $H_A : \mu \neq 3500$ and give a 95% confidence interval for μ. There are a few different ways to do this with Matlab. One way is to use the function `normfit`. The most convenient way, however, is to use the function tt ttest, which can answer all questions simultaneously. **Type**

$$[\mathtt{h}, \mathtt{p}, \mathtt{ki}] = \mathtt{ttest}(\mathtt{fv}, 3500, \mathtt{a})$$

and Matlab will give you

- `h:` rejection indicator, i.e. $h = 1$ means that the null hypothesis is rejected in favor of $H_A : \mu \neq 3500$ at the significance level a and $h = 0$ means that H_0 is accepted.
- `p:` The p-value of the test, i.e. the lowest possible significance level for which H_0 could have been rejected with this data set.
- `ki:` Upper and lower bounds for the symmetric confidence interval for μ with confidence level $1 - \alpha$.

You should arrive at a significance level of ca $2.2 \cdot 10^{-6}$, so the test is strongly significant, and a confidence interval of $\mu = 3400 \pm 41$. So the rule $\mu = 3500$ does not seem to be quite correct and the expectation rather seems to be close to 3400 g.

OK, but in this data set, one has recorded all the birth weights, including the children that were born early. One classes a child as born early if the length of the pregnancy was less than 266 days. *Now try* to test the same null hypothesis on the data set that excludes those early. Information about the length of the pregnancy is found in column 1 of `birth`. You will see that the result is quite different. If you also make a normal distribution probability plot of this data set, you will also see a better fit. Can you explain why this is the case?

Next, we move to the two sample case. Here we have two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ where $X_k \sim \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y_k \sim \mathrm{N}(\mu_y, \sigma_y^2)$ and we are usually interested in testing $H_0 : \mu_x = \mu_y$ against $H_A : \mu_x \neq \mu_y$. For an exact analysis, one usually has to assume that $\sigma_x^2 = \sigma_y^2 = \sigma^2$. In that case

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{s_P \sqrt{1/n + 1/m}} \sim t_{n+m-2},$$

where s_P^2 is the pooled sample variance given by

$$s_P^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

From this one infers e.g. the symmetric confidence interval

$$\mu_x - \mu_y = \bar{X} - \bar{Y} \pm F^{-1}_{t_{n+m-2}}(1 - \alpha/2) s_P \sqrt{\frac{1}{n} + \frac{1}{m}}$$

with confidence level $1 - \alpha$ and H_0 is tested correspondingly.

Matlab has the function `ttest2` to handle this situation. As for `ttest`, this function can answer all the interesting questions simultaneously. Use Matlab's help to find out how the functions works. (As you will see, the function can even handle the situation with $\sigma_x^2 \neq \sigma_y^2$ (numerically). **Now test** if there is any difference between expected birth weight between boys and girls. Do the same thing for children of smoking mothers versus non-smoking mothers and for mothers that live together with the father versus those that don't. Try some more hypotheses if you have the time. (Information about the infants gender, mother's smoking

habits and if the mother lives together with the father can be found in columns 2, 20 and 18 respectively.)

Remark. Recall the multiple testing problem: "in any data set there is always something that is significant". This is something we would have had to take into serious consideration if this had been a sharp situation. Let us therefore regard this exercise as a hypothesis generating pilot study.

Now turn back to the one sample case and consider the power of the test of $H_0 : \mu \neq \mu_0$ against $H_A : \mu \neq \mu_0$. Recall that test is based on the fact that if H_0 is true, then

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

The test rejects H_0 at level α if

$$\frac{|\overline{X} - \mu_0|}{s/\sqrt{n}} > F^{-1}_{t_{n-1}}(1 - \alpha/2).$$

For $\mu_1 \neq \mu_0$, the *power* of the test at μ_1, $g(\mu_1)$, is the probability that the test rejects H_0 if the true value of μ is μ_1. To compute this, we need to know the distribution of the test statistic $T := \sqrt{n}(\overline{X} - \mu_0)/s$ when $\mu = \mu_1$. If $\mu = \mu_0$, this was no problem, but if $\mu = \mu_1$ we get a so called *non-central* t-distribution. This is in itself not a big problem, since there are tables for this distribution and it is of course programmed into any decent statistical software. A more pressing problem is that the non-central t-distribution, apart from the number of degrees of freedom, also has the unknown variance σ^2 as a parameter. Hence, to apply it, one needs a good guess of what the variance is. Since power computations are usually needed before a study is conducted, in order to determine how large the study needs to be in order to have a good chance of detecting a certain effect, one cannot wait for the data to use that to estimate σ^2. Hence, one really cannot do anything better than to settle for a qualified guess of σ^2. Sometimes one is lucky enough that there have been earlier but similar studies made, from which one can make some conclusions. Otherwise an uninformed guess is necessary.

Remark. Apparently is is often hard to make power computations objective. Nevertheless, they are sometimes of utmost importance. Consider for example a medical company that wants to test a new drug to see if it has an effect that is better than placebo (or sometimes better than some other drug on the market). Say that we are about to make the Phase 3 study, where our drug is finally going to be tested for its effect on humans. First we decide on how large an effect needs to

be to be worth-wile to detect, the so called smallest effect of *clinical significance*. Then we want to make a power computation in order to find out how large the study needs to be in order to give us a, say, 90% chance of detecting an effect of this size (of course, if the true effect is larger than that, our chances will be even bigger). Since studies of this kind are extremely expensive, we definitely do not want to oversize the study. On the other hand, undersizing is not a good option either, since that would increase the risk that the study becomes just a big waste of money.

Now use the study of weights at birth for an example. If the true expected weight is 3450 g, then how many births would have had to be studied in order to get a 75% probability of rejecting $H_0 : \mu = 3500$ at the 5% significance level. As a guess of σ^2, use the rule of thumb that $\sigma = 500$ g. The Matlab function `sampsizepwr` will help you out. It can also be used to compute the power for given n (and μ_1, μ_0 and σ^2).

5 Linear regression

Recall that the setting where we used linear regression is when we have data in pairs, (x_k, Y_k), $k = 1, \ldots, n$, where the X_k:s are known quantities and the Y_k are assumed to depend linearly on the x_k:s, but with a normally distributed deviation, independent for different k but all with the same variance. In other words, there are constants a and b (and σ^2) such that $Y_1, \ldots, Y_n$ are independent and

$$Y_k \sim \mathrm{N}(a + bx_k, \sigma^2).$$

An equivalent way of writing this is that $Y_k = a + bx_k + \epsilon_k$, where $\epsilon_1, \ldots, \epsilon_n$ are independent and $\mathrm{N}(0, \sigma^2)$ distributed. From the observations, one wants to estimate the unknown parameters a and b (and σ^2). Using the ML method, it turns out that one finds the point estimators $\hat{a}$ and $\hat{b}$ of a and b, by minimizing

$$\sum_{k=1}^{n} (Y_k - (a + bx_k))^2,$$

over a and b. In other words, the ML method coincides with the method of least squares for approximating the over-determined system of equations $Y_k = a + bx_k$, $k = 1, \ldots, n$ (regarding a and b as the variables and the x_k:s and Y_k:s as known coefficients). One also gets that $\hat{a}$ and $\hat{b}$ get normal distributions whose variances

depend on σ^2, which in turn leads to that one can make tests and confidence intervals for a and b, using the t-distribution (or the normal distribution if σ^2 is known).

The easiest way to get $\hat{a}$ and $\hat{b}$ from Matlab is to use `polyfit(x,Y,1)`, where of course `x` and `Y` are the vectors of the x_k:s and the Y_k:s respectively. To visualize this, use `scatter(x,Y)` or `plot(x,Y,'.')` (which I think looks better) and give the command `lsline` to plot the regression line through the data points. To get confidence intervals for a and b is not as straightforward. The function to use is `regress`. This function is based on the *general linear model* (GLM), which we have not been treating in this course. What we need to know here, is that linear can be seen as a special case of GLM, by writing $Y_k = a + bx_k + \epsilon_k$, $k = 1, \ldots, n$, on matrix form as

$$Y = C\beta + \epsilon$$

where $Y = [Y_1 \ldots Y_n]^T$, $\beta = [a\ b]^T$ and C is the $n \times 2$ matrix whose columns are $[1 \ldots 1]^T$ and $[x_1 \ldots x_n]^T$. The function `regress` has Y and C as its input and you create C by typing `C=[ones(n,1)  x]`. **Now use** this to make a linear regression of the weights at birth as a function of length of pregnancy (which you find in the first column of `birth`). What is the 99% confidence interval for b? You should get $\hat{b} \approx 28$. How do you interpret this?

6 Analysis of Variance

Analysis of Variance, ANOVA for short, is a generalization of the usual two sample situation to a situation with more than two samples. The model is the following. We have k independent samples

$$X_{11}, X_{12}, \ldots, X_{1n_1},$$

$$X_{21}, X_{22}, \ldots, X_{2n_2},$$

$$\vdots$$

$$X_{k1}, X_{k2}, \ldots, X_{kn_k}$$

where sample no. i, i.e. $X_{i1}, \ldots, X_{in_i}$, is taken from a $\mathrm{N}(\mu_i, \sigma^2)$-distribution. Thus we assume that the variance is the same for all the samples. We want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_n$ against H_A : not all μ_i equal.

Example. Consider an apartment building where one wants to investigate the radon level in the apartments, to see if there is an effect on the radon level from what floor the apartments are on; such an effect would indicate that one sours of radon is from the ground. On each level, one places a number of detectors to measure the radon levels and use the data to do an ANOVA.

One could of course use the two sample methodology for each pair of samples, but the one would then encounter the problem of multiple testing. Properly taking care of that problem would give a very inefficient test if k is large. In the present situation, we have a fairly weak alternative hypothesis, which only claims that *some* μ_i is different from the others. As we will see, this can be tested in a far more efficient way. The commonly used test statistic builds on the idea that if observed values varies more *between* the samples than *within* the samples, then this indicates that H_0 is false.

Let $n = \sum_{i=1}^{k} n_i$ be the total number of observations. Write

$$\overline{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

for the mean of the i'th sample and $\overline{X}$ for the mean of all observations, i.e.

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{X}_{i.}$$

Write also

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2$$

for the sample variance of the i'th sample. By Proposition 1.5, $(n_i - 1)s_i^2 \sim \chi^2_{n-1}$, so by summing

$$\sum_{k=1}^{n} \frac{(n_i - 1)s_i^2}{\sigma^2} \sim \chi^2_{n-k}.$$

By writing

$$s_W^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)s_i^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2$$

this transforms into

$$\frac{(n-k)s_W^2}{\sigma^2} \sim \chi^2_{n-k}.$$

Here the index W stands for "within samples" to signify that s_W^2 is to be thought of as the in-sample variation of data.

Moving on, we have that $\overline{X}_{i.} \sim \mathrm{N}(\mu_i, \sigma^2/n_i)$. To make things easier for a while, assume that all the samples are of equal size, i.e. there is an m so that $n_i = m$ for all i. If the null hypothesis is true, then there is also a μ such that $\mu_i = \mu$ for all i. Hence $\overline{X}_{1.}, \overline{X}_{2.}, \ldots, \overline{X}_{k.}$ is a sample from the $\mathrm{N}(\mu, \sigma^2/m)$ distribution. Therefore

$$\frac{\sum_{i=1}^k m(\overline{X}_{i.} - \overline{X})^2}{\sigma^2} \sim \chi^2_{k-1}.$$

Writing

$$s_B^2 = \frac{1}{k-1} \sum_{i=1}^k m(\overline{X}_{i.} - \overline{X})^2,$$

this becomes

$$\frac{(k-1)s_B^2}{\sigma_2} \sim \chi^2_{k-1}.$$

The index B stands for "between samples" to signify that s_B^2 is thought of as the between-sample variation of data. Returning to the general case of (possibly) different n_i:s, we take

$$s_B^2 = \frac{1}{k-1} \sum_{i=1}^k n_i(\bar{X}_{i.} - \bar{X})^2.$$

Again it can be shown (but we will not do that here) that $(k-1)s_B^2/\sigma^2 \sim \chi^2_{k-1}$. One can also show that s_W^2 and s_B^2 are independent. This means that the ratio $T := s_B^2/s_W^2$ has an *F-distribution.*

Definition 6.1 *Let Y_1 and Y_2 be two independent random variables with $\chi^2_{m_1}$ and $\chi^2_{m_2}$ distributions respectively. Then the ratio*

$$\frac{Y_1/m_1}{Y_2/m_2}$$

is said to have an F-distribution with m_1 and m_2 degrees of freedom. One writes for short

$$\frac{Y_1/m_1}{Y_2/m_2} \sim F_{m_1,m_2}.$$

It follows immediately from the definition that if H_0 is true, then

$$T = \frac{s_B^2}{s_W^2} \sim F_{k-1,n-k}.$$

If T becomes large, then this speaks in favor of H_A over H_0. Hence the test for testing H_0: "all μ_i equal" against H_A : "not all μ_i equal" is given by rejecting H_0 on the significance level α if

$$T \geq F^{-1}_{F_{k-1,n-k}}(1-\alpha).$$

Exercise: A small bus company has only five buses, all of the same type. The company wants to try out four different types of tires with respect to wear. An experiment is conducted, where one tyre of each type is put on each of the five buses. After 20000 km of use, the wear of each tyre is measured (in mm) and recorded. This gives a sample $X_{i1}, \ldots, X_{i5}$ for each tyre type i, $i = 1, 2, 3, 4$. The data is given by

Tire type 1: 9.1 13.4 15.6 11.0 17.1

Tire type 2: 20.3 20.3 24.6 18.2 19.8

Tire type 3 3: 20.8 28.3 23.7 21.4 25.1

Tire type 4: 11.8 16.0 16.2 14.1 15.8

May we assume that data is normal? Making a normal probability plot for each sample looks OK, so let assume that we can. Do the samples have the same variance? Computing the sample variances, this also looks reasonable. Are the samples independent? This also seems reasonable since the different buses are of the same type. Then it is at least not too unreasonable to assume the the variation in wear depends on the tires themselves rather than a variation between different buses. Hence it seems to be OK to do an ANOVA (but we should be aware that the samples are small, so it hard to check our assumptions, which of course makes them quite uncertain).

The purpose of the experiment seems to be to test $H_0 : \mu_1 = \ldots = \mu_5$ against H_A : "not all μ_i equal. We have a *balanced* experiment, i.e. all n_i are equal. In such a case the Matlab function `anova1` is very convenient. Matlab will answer with the p-value of the test (and a figure and a table that you don't need to consider). If you want, you can compare p-values with p-values for the corresponding

pairwise two sample t-tests (without thinking too deeply about what the comparison means).

Remark. The "`1`" in `anova1` stands for that we are making a *one-way* ANOVA. There is also two-way (and n-way) ANOVA, a slightly different model that we will not be concerned with here.

Exercise: In the matrix `birth` with birth weight observations, there is information about the age of the mother at the time of birth. This is in the form of "1" if the mother was 15-24 years, "2" if she was 25-29 years and "3" if she was at least 30 years. This information is found in the eighth column. Now partition the observations into three groups based on the age of the mother and do an ANOVA to test if there is an age effect on the birth weights.

You will find that this is an unbalanced experiment. Matlab does not have any good function for this, so you have to compute the value of the F-distributed test statistic T and insert into the distribution function for the correct F-distribution. The distribution function is implemented in Matlab: `fcdf`.

7 Non-parametric methods

In some situations one cannot defend any assumption on the distribution of data. There are still things that one can do. We will see a few examples of this here. Assume that $X_1, \ldots, X_n$ is a sample from a distribution about which we know nothing more than that it is continuous.

Definition 7.1 *Let X be random variable with distribution function F. Then the number m is called a median for X (or for F) if $F(m) = 1/2$.*

The definition applies to any random variable, but if F is not continuous, then there may be no median. Since our X under consideration is assumed to be continuous, we have that F is continuous and hence a median must exist (by the Intermediate Value Theorem). The median may not be unique, but if F is strictly increasing uniqueness always holds. Note that the median is *not* the same thing as the expectation, even if it exists and is unique. If the distribution of X is *symmetric* around m, i.e. $f(m+x) = f(m-x)$ for all x, then $m = \mathbb{E}[X]$, but not in general. Consider for example $X \sim \exp(1)$. Then $\mathbb{E}[X] = 1$, but $P(X > x) = e^{-x}$ which is $1/2$ for $x = \ln 2$, so $m = \ln 2$.

For simplicity, assume that X has a unique median (even though everything we will do works without this assumption). The median m is thus a number such

that $\mathbb{P}(X < m) = \mathbb{P}(X > m) = 1/2$. Fix a number m_0. Let N_+ be the number of observations X_k such that $X_k > m_0$. Now *if* $m = m_0$, then N_+ has a binomial distribution with parameters n and $1/2$ and we should get N_+ close to $n/2$ and a large deviation will indicate that $m \neq 1/2$. The stage is set for a test. Let c be such that

$$F_{\mathrm{Bin}(n,1/2)}(n/2 + c) \geq 1 - \alpha/2$$

and reject $H_0 : m = m_0$ in favor of $H_A : m \neq m_0$ at significance level at most α if

$$|N_+ - n/2| > c.$$

(That we typically cannot get exactly significance level α is because the binomial distribution is discrete.) The test is called a *sign test* of the median. The corresponding Matlab function is `signtest`. **Try this** on the birth data to test if the median is 3400 g.

The fact that $N_+ \sim \mathrm{Bin}(n, 1/2)$ if $m = m_0$ (or if we replace m_0 with m in the definition of N_+) can of course also be used to make confidence intervals for m.

Remark. The vector `fv` of birth weights, has a sample mean of 3400 g and a sample median of 3430 g. We have assumed that data are normal and since the normal distribution is symmetric about its mean, the mean and the median coincide. Hence we should expect the sampled quantities to be very close. However, we did observe that that there are to many very low weights for a really good fit with the normal distribution. In the light of this, it is not surprising that the sample median is a bit larger than the sample mean. If we instead restrict to birth weights for children not born early, we have a much better fit with normality and in this case the sample median is 3480 g and the mean 3496 g, i.e. much closer.

Observe that the sign test only takes into consideration the number of observations above m_0 and not how far above m_0 they are compared to how far below the other observations are. This is as it should be, since the median m itself does only take into consideration the amount of probability mass to the left/right of m and not where it is. However, suppose that we have good reason to believe that data comes from a symmetric distribution. Then the mean and the median coincide, so testing $H_0 : m = m_0$ is the same as testing $\mu = \mu_0$. The mean however, does very much depend on where the probability mass is. Hence it does in this situation make good sense to take the position of the observations into account. We do this in the following way. Rank the numbers $|X_k - \mu_0|$ from the smallest to the largest. For each k, let $R_k = j$ if $|X_k - \mu_0|$ is the j:th smallest of these numbers. In this way every observation X_k gets its *rank* R_k in terms of how much it deviates from

μ_0. Next, define the test statistic

$$W := \sum_{k:X_k>\mu_0} R_k$$

i.e. the rank sum of the observations that exceed μ_0. This test statistic will detect if the observations on one side of μ_0 deviate more from μ_0 than the ones on the other side.

The range of W is $0, 1, \ldots, n(n+1)/2$. If μ_0 is the true expectation, then it is fairly easy to see that $\mathbb{E}[W] = n(n+1)/4$ and that W:s pdf is symmetric around this value. More precisely, one can show that if μ_0 is the true mean, then

$$\mathbb{P}(W = r) = \frac{a(n)}{2^r},$$

where $a(r)$ is the coefficient for s^r in the expansion of $\prod_{j=1}^{n}(1+s^j)$. Let F be the distribution function for a random variable with this distribution. Find c so that $F(n(n+1)/4 + c) \geq 1 - \alpha/2$ and reject $H_0 : \mu = \mu_0$ in favor of $H_A : \mu \neq \mu_0$ at level at most α if

$$\left| W - \frac{n(n+1)}{4} \right| > c.$$

This test is called the *Wilcoxon signed rank test* (WSignrank). The coefficients $a(r)$ are difficult to express explicitly. For n large, this is is largely overcome by the fact that, if H_0 is true, then

$$\frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \approx \mathrm{N}(0,1).$$

In any case, with access to Matlab, computing the $a(r)$:s is not a problem. The Matlab function for WSignrank is `signrank`. **Check out** how it works and try it on the same birth weight data that you did with the sign test. You should find that corresponding p-values are smaller for WSignrank than for the sign test. Hence WSignrank is more efficient. On the other hand it requires a symmetry assumption that the sign test does not need. Compare also WSignrank with the corresponding t-test under the assumption that data are normal. Here you should find that the t-test is usually, but not always, more efficient.

Next we consider non-parametric comparison of two samples. We will assume a *translation model*. We have the two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and assume that they come from distributions with distribution functions F_1 and F_2 respectively, where these have the same form but may differ in

that one is a translate of the other, i.e. there exists t such that $F_1(x+t) = F_2(x)$ for all x.

We wish to test $H_0 : F_1 = F_2$ (which by assumption is the same as testing $H_0 : \mu_1 = \mu_2$ or $H_0 : t = 0$) against $H_A : F_1 \neq F_2$. Assume without loss of generality that $m \leq n$. If the null hypothesis holds, then we may regard the whole collection of data, $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ as a single sample of size $m+n$ from the common distribution. Consequently, if one ranks the $m+n$ observations by size, all $(m+n)!$ possible rankings are equally probable. Now let the test statistic be given by

$$W := \sum_{k=1}^{m} r(X_k),$$

where $r(X_k)$ is the rank of X_k in the unified sample. If H_0 is true, then the expected rank of one observation os $(m+n+1)/2$, so $\mathbb{E}[W] = m(m+n+1)/2$ and the distribution of W is symmetric around this number. Let F be the distribution function of W under H_0. In analogy with the above, choose c so that $F(m(m+n+1)/2 + c) \geq 1 - \alpha/2$ and reject $H_0 : F_1 = F_2$ in favor of $H_A : F_1 \neq F_2$ at significance level at most α if

$$\left| W - \frac{m(m+n+1)}{2} \right| > c.$$

This test is called the *Wilcoxon rank sum test* (WRanksum). To compute F explicitly is difficult, but not a problem for Matlab. There is a central limit theorem here too, see the book at page 373. The Matlab function for the test is `ranksum`. **Now make** a rank sum test to see if birth weights differ between children of smoking mothers and children of non-smoking mothers. Compare with the two sample t-test you did before.

Discussion. Generally speaking, non-parametric models are better than parametric models (e.g. models based on the normal distribution) in that they make very few assumptions on the distribution of data. On the other hand, the parametric models are usually stronger, i.e. they have a greater chance of detecting a given deviation from the null hypothesis. Thus they both have essential advantages and they both have a central rôle in applied statistics.