

ASTR 3890 - Selected Topics: Data Science for Large
Astronomical Surveys (Spring 2022)

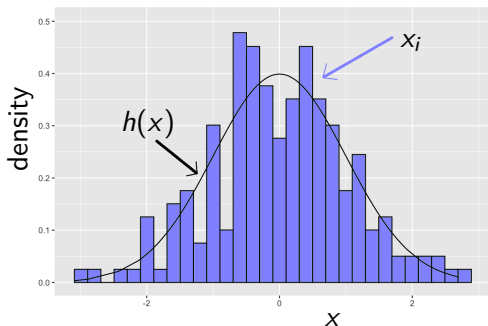
Introduction To Probability & Statistics: II

Dr. Nina Hernitschek
February 14, 2022

recap: Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i , not the underlying pdf



recap

Descriptive
Statistics

Sample versus
Population
Statistics

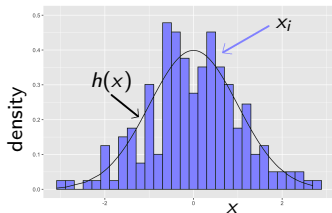
Distributions

Transformations
of Random
Variables

recap: Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i , not the underlying pdf

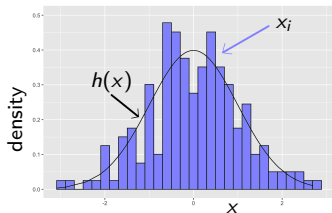


using measurements x_i , we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn

recap: Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i , not the underlying pdf



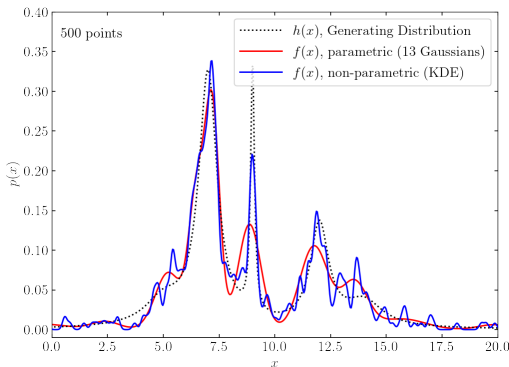
Question: What is the probability of a value lying between x and dx ?

recap: Physical Models

idea: We can either:

- Describe the data, then the process is non-parametric, i.e. we are just trying to **describe** the data behavior in a compact practical way.
- Guess a physical model for $h(x)$, then the process is parametric.

From a **model** we can generate new data that mimic measurements.



Descriptive Statistics

Our goal is to estimate $h(x)$ given some measured data, by reconstructing the data-based distribution $f(x)$.

An arbitrary distribution* can be characterized by location parameters (i.e., position), scale parameters (i.e., width), and shape parameters. These parameters are called **descriptive statistics**.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

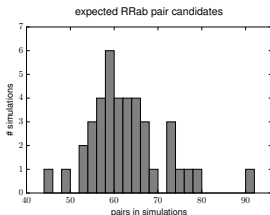
Transformations
of Random
Variables

Descriptive Statistics

Our goal is to estimate $h(x)$ given some measured data, by reconstructing the data-based distribution $f(x)$.

An arbitrary distribution* can be characterized by location parameters (i.e., position), scale parameters (i.e., width), and shape parameters. These parameters are called **descriptive statistics**.

*can be anything, e.g. the distribution of velocities in a globular cluster, the distribution of the number of epochs in light curves...



Descriptive Statistics

mean of a sample:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

This is the **sample arithmetic mean**

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

mean of a sample:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

This is the **sample arithmetic mean**

which is derived as the first momentum of the distribution from Monte Carlo integration as

$$\mu = E(x) = \langle x \rangle = \int_{-\infty}^{+\infty} x h(x) d(x) \approx \frac{1}{N} \sum_{i=1}^N x_i$$

where $\{x_i\}$ are random samples from the properly normalized $h(x)$, and $E(\cdot)$ is the **expectation value**

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$

if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,

if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$

example:

1 3 3 4 7 8 128

if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,


if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$

if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

example:

1 3 3 **4** 7 8 128

median = 4

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,


if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$

if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

example:

1 3 3 **4** 7 8 128

median = 4

example:

1 3 3 4 7 8

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

median of a sample:

The **median** is a more robust estimator than the mean (see Jupyter notebook `lecture_3.ipynb`)

The median is defined as follows: For a distribution x with n elements, ordered from smallest to greatest,


if n is odd,

$$\text{median}(x) = x_{(n+1)/2}$$


if n is even,

$$\text{median}(x) = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

example:

1 3 3 **4** 7 8 128

median = 4

example:

1 3 **3** **4** 7 8

median = 3.5

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

“average” location value (the mean, median) discussed before
⇒ information about *deviations* from the average (which is related to the shape of the distribution)

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

“average” location value (the mean, median) discussed before
⇒ information about *deviations* from the average (which is related to the shape of the distribution)

One could start with a deviation such as

$$d_i = x_i - \mu$$

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

“average” location value (the mean, median) discussed before
⇒ information about *deviations* from the average (which is related to the shape of the distribution)

One could start with a deviation such as

$$d_i = x_i - \mu$$

However, the average deviation is zero by definition of the mean.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

One can compute the **mean absolute deviation (MAD)**:

$$\frac{1}{N} \sum |x_i - \mu|$$

but the absolute values can hide the true scatter of the distribution in some cases

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

Better idea: square the differences

$$\sigma^2 = \frac{1}{N} \sum (|x_i - \mu|)^2$$

which is the **variance**.

The variance V is the expectation value of $(x - \mu)^2$ (and related to the 2nd moment)

$$\sigma^2 = V = E((x - \mu)^2) = \int_{-\infty}^{+\infty} (x - \mu)^2 h(x) dx$$

where σ is the **standard deviation**.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

Other descriptive statistics are related to **higher order moments** of the distribution:

Better idea: square the differences

$$\sigma^2 = \frac{1}{N} \sum (|x_i - \mu|)^2$$

which is the **variance**.

The variance V is the expectation value of $(x - \mu)^2$ (and related to the 2nd moment)

$$\sigma^2 = V = E((x - \mu)^2) = \int_{-\infty}^{+\infty} (x - \mu)^2 h(x) dx$$

where σ is the **standard deviation**.

For discrete distributions, the integral gets replaced by a sum.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

The **Median Absolute Deviation** (also MAD) given by

$$\text{median}(|x_i - \text{median}(\{x_i\})|)$$

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

In statistics and probability, **quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

A **quartile** is a quantile which divides the distribution into four parts of equal size.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Descriptive Statistics

recap

Descriptive
Statistics

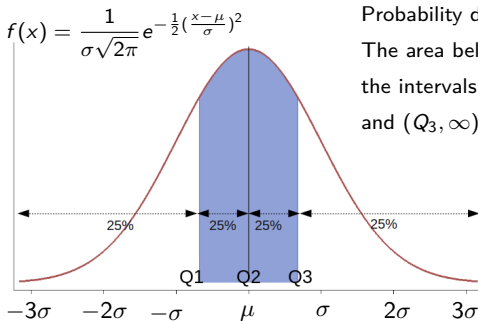
Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

In statistics and probability, **quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.

A **quartile** is a quantile which divides the distribution into four parts of equal size.



Probability density of a normal distribution.

The area below the red curve is the same in the intervals $(-\infty, Q_1)$, (Q_1, Q_2) , (Q_2, Q_3) , and (Q_3, ∞) .

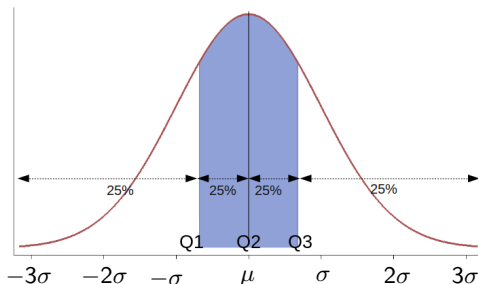
Descriptive Statistics

P % quantiles (or the p^{th} **percentile**, q_p) are computed as

$$\frac{p}{100} = H(q_p) = \int_{-\infty}^{q_p} h(x) dx$$

The full integral from $-\infty$ to ∞ is 1 (100%). So, here you are looking for the value of x that accounts for P percent of the distribution.

For example, the 25th, 50th, and 75th percentiles are just Q1, Q2, Q3



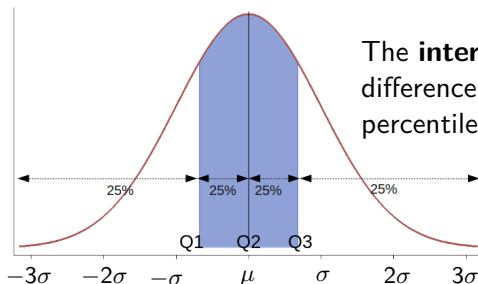
Descriptive Statistics

P % quantiles (or the p^{th} **percentile**, q_p) are computed as

$$\frac{p}{100} = H(q_p) = \int_{-\infty}^{q_p} h(x) dx$$

The full integral from $-\infty$ to ∞ is 1 (100%). So, here you are looking for the value of x that accounts for P percent of the distribution.

For example, the 25th, 50th, and 75th percentiles are just Q1, Q2, Q3



The **interquartile range** is the difference between the 25th and 75th percentiles, $q_{75} - q_{25}$

Descriptive Statistics

The **mode** is the most probable value, determined from the peak of the distribution, which is the value where the derivative is 0 (i.e. the turning point):

$$\left(\frac{dh(x)}{d(x)} \right)_{x_m} = 0$$

For a Gaussian distribution, the mode is its mean μ .

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

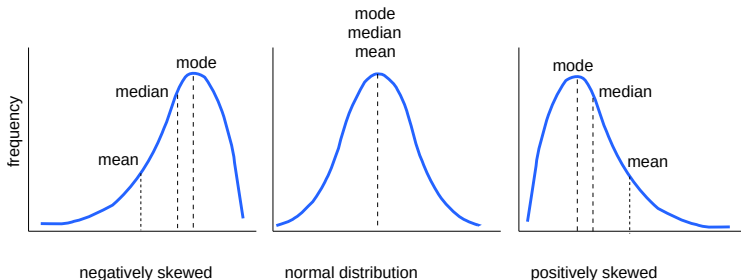
Descriptive Statistics

Other useful shape measures include the “higher order” moments:

Skewness:

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^3 h(x) dx$$

The skewness measures the distribution's asymmetry. Distributions with long tails to larger x values have positive Σ .



Descriptive Statistics

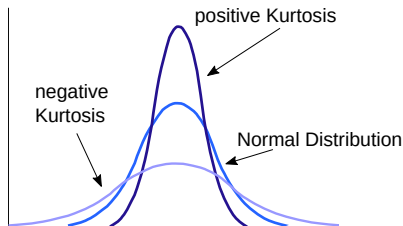
Other useful shape measures include the “higher order” moments:

Kurtosis:

$$K = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^4 h(x) dx - 3$$

The kurtosis measures how peaked or flat-topped a distribution is, with strongly peaked ones being positive and flat-topped ones being negative.

K is calibrated to a Gaussian distribution (hence the subtraction of 3).

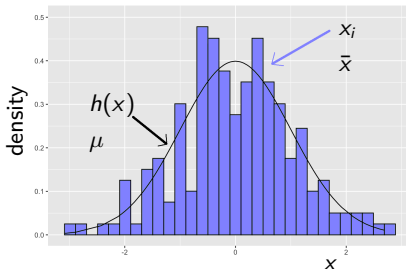


Sample versus Population Statistics

Statistics estimated from the data are called **sample statistics** as compared to **population statistics** derived from knowing the functional form of the pdf.

Specifically, μ is the **population mean**, i.e. it is the expectation value of x for $h(x)$. But we don't know $h(x)$. So the **sample mean** \bar{x} is an estimator for μ . The sample mean is defined as

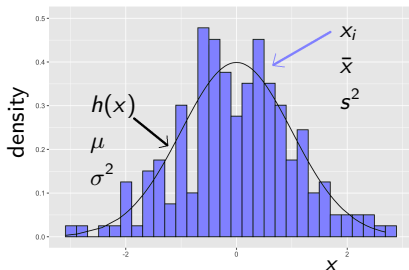
$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$



Sample versus Population Statistics

Instead of the **population variance** σ^2 , we have the **sample variance** s^2 , where

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$



The denominator $N - 1$ (instead of N) accounts for the fact that we determine \bar{x} from the data itself instead of using a known μ . Ideally one tries to work in a regime where N is large enough to ignore this.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Uncertainty of Sample Statistics

What are the uncertainties of our estimates \bar{x} and s^2 ?

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Uncertainty of Sample Statistics

What are the uncertainties of our estimates \bar{x} and s^2 ?

Note that s is the width estimate of the underlying distribution; it is not the uncertainty of \bar{x} .

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Uncertainty of Sample Statistics

What are the uncertainties of our estimates \bar{x} and s^2 ?

Note that s is the width estimate of the underlying distribution; it is not the uncertainty of \bar{x} .

The uncertainty of \bar{x} , $\sigma_{\bar{x}}$, is

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

which we call the **standard error of the mean**. The uncertainty of s itself is

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}} \sqrt{\frac{N}{N-1}} \sigma_{\bar{x}}$$

Note that for large N , $\sigma_{\bar{x}} \sim \sqrt{2}\sigma_s$, and for small N , σ_s is not much smaller than s .

Distributions

A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Distributions

A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

What's the point of all these distributions?

To understand the **significance** of a measurement, we want to know how likely it is that we would get that measurement in our experiment by random chance. To determine that we need to know the shape of the distribution.

If we are attempting to characterize our data in a way that is **parameterized**, then we need a functional form for a distribution.

recap

Descriptive
Statistics

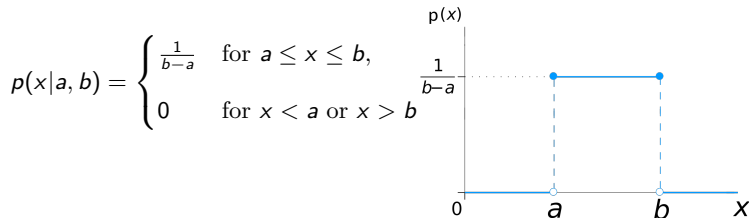
Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Uniform Distribution

The probability density function of the **continuous uniform distribution** is given by:



The mean (first moment) of the distribution is:

$$E(X) = \frac{1}{2}(b + a)$$

The variance (second central moment) is:

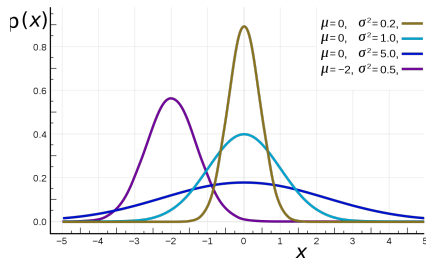
$$V(X) = \frac{1}{12}(b - a)^2$$

Gaussian Distribution

The probability density function of the **Gaussian distribution** is given by:

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right)$$

It is also called the **normal distribution** and can be noted by $\mathcal{N}(\mu, \sigma)$.



recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables

Gaussian Distribution

Gaussian confidence levels

The probability of a measurement drawn from a Gaussian distribution that is between $\mu - a$ and $\mu + b$ is

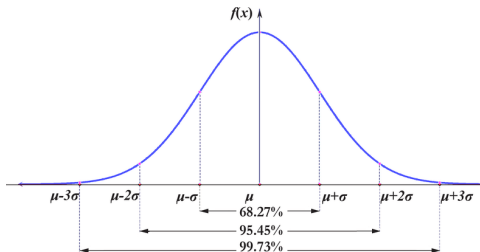
$$\int_{\mu-a}^{\mu+b} p(x|\mu, \sigma) dx$$

examples:

for $a = b = 1\sigma$, we get 68.3%

for $a = b = 2\sigma$, we get 95.4%

for $a = b = 3\sigma$, we get 99.7%



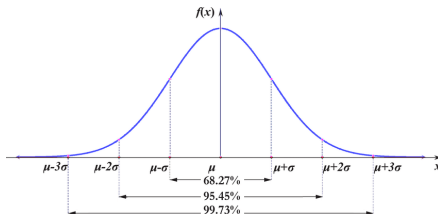
We refer to the ranges $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ as the 68%, 95% and 99% **confidence limits**, respectively. Note: These numbers are only valid for Gaussian distributions (but can be calculated for other distributions).

Gaussian Distribution

Gaussian confidence levels

The probability of a measurement drawn from a Gaussian distribution that is between $\mu - a$ and $\mu + b$ is

$$\int_{\mu-a}^{\mu+b} p(x|\mu, \sigma) dx$$



confidence level vs. confidence interval:

The **confidence level** is the percentage of all possible samples that are expected to include the true population parameter: the percentage of times you expect to get close to the same estimate if you repeat the experiment.

The **confidence interval** is the upper and lower bounds of the estimate you expect to find at a given level of confidence - an interval that is likely to contain an unknown population parameter.

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

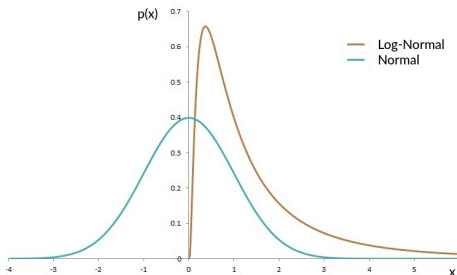
Transformations
of Random
Variables

Log-Normal Distribution

If x is Gaussian distributed with $\mathcal{N}(\mu, \sigma)$, then $y = \exp(x)$ has a **log-normal distribution**, where the mean of y is $\exp(\mu + \sigma^2/2)$, the median is $\exp(\mu)$ and the mode is $\exp(\mu - \sigma^2)$.

The probability distribution function for a log-normal distribution is

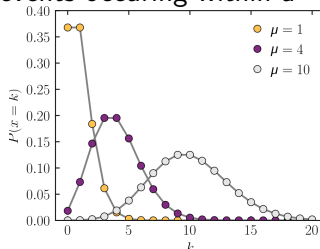
$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$



Poisson Distribution

The **Poisson distribution** is a distribution for a discrete variable, telling the probability of k events occurring within a certain time when the mean is μ :

$$p(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!}$$



The mean μ completely characterizes the distribution. The mode is $(\mu - 1)$, the standard deviation is $\sqrt{\mu}$.

As μ increases, the Poisson distribution becomes more and more similar to a Gaussian with $\mathcal{N}(\mu, \sqrt{\mu})$.

In the plot, the horizontal axis k , the number of occurrences. μ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given μ . The function is defined only at integer values of k .

Transformations of Random Variables

If x is a random variable then $f(x)$ is also a random variable for any function f . To transform probability distributions when taking functions of random variables, we can simply use conservation of dimensionless probability, i.e.

$$\text{Prob}(x, x + dx) = \text{Prob}(y, y + y)$$

$$\Rightarrow p(x)dx = p(y)dy \text{ with } y = f(x)$$

Thus

$$p(y) = |dx/dy|p(x)$$

Example:

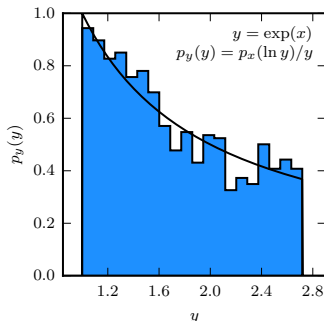
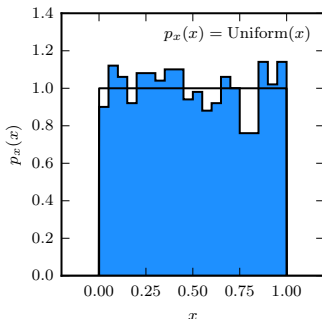
Let x be a random variable drawn from a uniform distribution between 0 and 1. So $p(x) = 1/(1 - 0) = 1$. Let's transform to

$$y = e^x: p(y) = \left| \frac{dy}{dx} \right|^{-1} p(x) = 1/y$$

Transformations of Random Variables

Example:

Let x be a random variable drawn from a uniform distribution between 0 and 1. So $p(x) = 1/(1 - 0) = 1$. Let's transform to $y = e^x$: $p(y) = \left| \frac{dy}{dx} \right|^{-1} p(x) = 1/y$



Break & Questions

afterwards we continue with `lecture_4.ipynb` from the github repository

recap

Descriptive
Statistics

Sample versus
Population
Statistics

Distributions

Transformations
of Random
Variables