TRISEP - 2022

Reference notes for Statistics

D. Karlen / University of Victoria and TRIUMF

Table of contents

Probability Theory

D. Karlen / University of Victoria and TRIUMF

Probability theory

- The term "probability" arises in situations where we lack complete certainty
  - Mathematics – certainty (prove theorems are true or false)
  - Experimental sciences – uncertainty
    - the goal of experimental science is to improve our understanding of things (the Universe, atoms, rabbits, …)
    - progress is made by making observations that reduce our uncertainty
    - a state of absolute certainty is seldom reached
      - not generally realized by popular media and general public
  - Statistics – the application of probability theory to experimental science

Meaning of probability

- The word "probability" is common in everyday language and, for many, it has an intuitive meaning that is not easily expressed
  - scientists need to be careful to use the term in a self-consistent manner

- Question: Does "probability" have the same meaning in the two following sentences?
  - The probability that a flipped coin shows heads is 50%.
  - The probability that it will rain tomorrow is 50%.

Meaning of probability

- Question: Is "probability" a property of a object/system?

  - Example: Suppose Jane says the following:
    - "This open, black bottle contains 10 marbles: 2 are red and 8 are blue. The mass, m, of the bottle is 1kg. If I shake the bottle and it is tipped on its side, the probability, p , that a red marble is the first to come out is 20%."

  - It seems clear that the mass, m, is a property of the bottle
  - One could argue that the probability, p, is
    - a property of the bottle (analogous to the mass); or
    - a property of Jane, expressing her degree of certainty of a future outcome of shaking and tipping the bottle

Two approaches

- Two approaches to probability theory have developed using these two different interpretations of probability

  - Original approach
    - now known as "Bayesian" or "subjective"
    - probability refers to the state of knowledge about a system – it is a not a property of the system itself
    - probability = degree of belief

  - New (19th century) approach
    - known as "frequentist", "classical", or "orthodox"!
    - probability refers to a property of the system
    - probability = relative frequency of occurrence

Which approach to follow?

- Most elementary data analysis textbooks for science students follow the frequentist approach, and do not even acknowledge that another approach exists
  - ad hoc recipes and rules are given
  - contradictions and paradoxes arise when other the other form of probability is considered

- Practicing scientists usually follow the frequentist approach in publishing results from experiments
  - most feel that it is too subjective to consider probability as a state of knowledge
    - it requires one to assume a particular state of knowledge before the experiment, in order to describe the state of knowledge after an experiment

Which approach to follow?

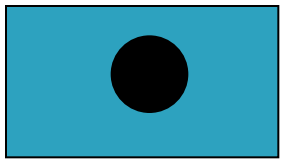
- Frequentist approach cannot answer the important questions, unlike the Bayesian approach
  - eg. How certain are you that this is a new discovery?

- Because of current practice, I will focus on the frequentist approach…
  - it is important to be able to converse with the majority of scientists

- I will point out the pitfalls along the way
  - equally important to appreciate the limitations of frequentist methods

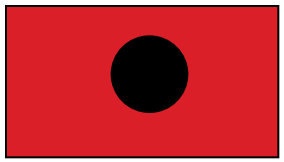
Frequentist probability: randomness

- A system which behaves in a manner that cannot be predicted with complete certainty is said to be random.
  - Consider the two pushbutton boxes below. Which one might be random?



1 1 1 1 1 1 1 1 1 1

ejs_buttons.jar



1 0 1 1 0 1 0 1 1 1

- With the given information, we cannot be certain if either box produces data with a predictable pattern.
- When we define a model to describe the boxes, we can choose to characterize some aspects of it to be random.

Random variables

- Suppose we characterize the red box as random
- use the symbol "*b*" to represent an observed outcome of pushing the button:
  - *b* is either 0 or 1
  - it has a known value – it should not be considered to be "random"
- use the symbol "*B*" to represent a possible (or future) outcome of pushing the button
  - *B* is an unknown value – it is called a "random variable"
    - Random variables are used to represent data NOT YET OBSERVED
  - although *B* is an unknown value, there is other information about *B* that we may know
    - Example: the probability for an outcome of *B* to be 1 may be known.

Frequentist probability

- Frequentists generally consider probability to be a property of the random system
- The probability P(B=1) is defined as:

$$P(B=1) = \lim_{n \to \infty} \frac{\text{number of occurrences of } b=1 \text{ in } n \text{ trials}}{n}$$

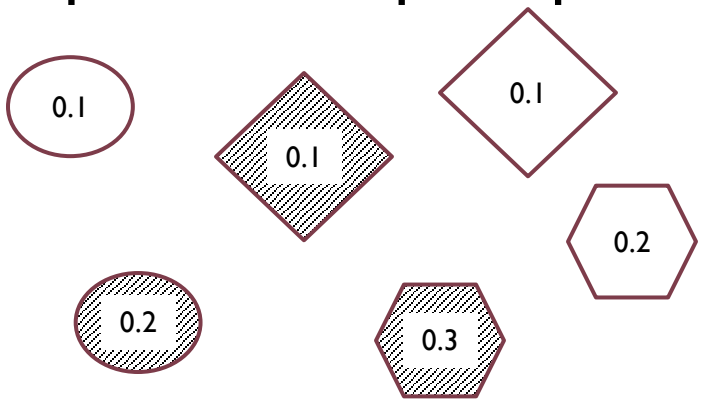
  - A circularity problem exists in this definition, if you try to put it into practice:
    - Repeated operations are needed to produce the trials
      - what is required in these repeated operations – exactly the same conditions? Similar conditions? How you decide what conditions?
      - If the condition is that the probability is unchanged, then the definition is circular.

A consistent framework for frequentists

- A consistent frequentist should understand that:
  - In all statistical analyses, a mathematical model of the experiment is necessary to interpret the data.
  - Randomness is a property of the model. One does not have to assert that the physical system is random.
  - Random variables are used only to refer to outcomes of the model, not data from real experiments.
    - The probability definition is therefore realizable, it can refer to model trials repeated under exactly the same conditions.
  - Any statements about probability may only refer to outcomes of the model.
  - Interpretations about the physical system following the experiment assumes that the model is valid

Probability axioms

- Generally accepted rules for probability:
  - $0 \leq P(B = i) \leq 1$
    - where i is an outcome (aka "event")
  - $P(B = i \text{ or } B = j) = P(B = i) + P(B = j)$
    - provided that i and j are mutually exclusive
  - $P(B \text{ in } S) = 1$
    - where S is the space of all possible outcomes
- Graphical example – probability for shapes & shades:

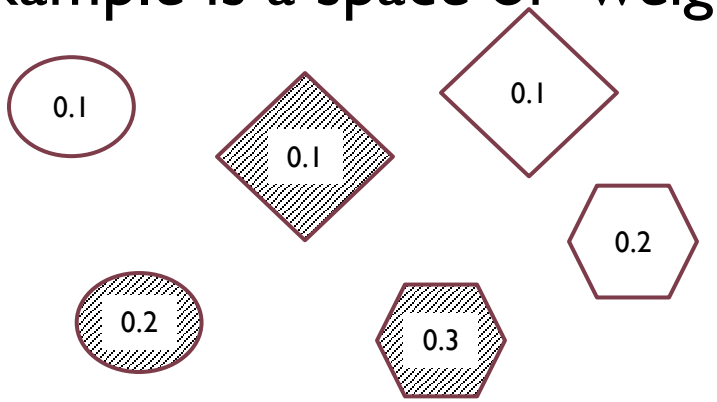


- $P(\text{shaded oval}) = 0.2$
- $P(\text{oval}) = 0.1 + 0.2 = 0.3$
- $P(\text{shaded}) = 0.6$
- $P(\text{oval or hexagon}) = 0.8$
- $P(\text{oval or shaded}) \neq 0.3 + 0.6$

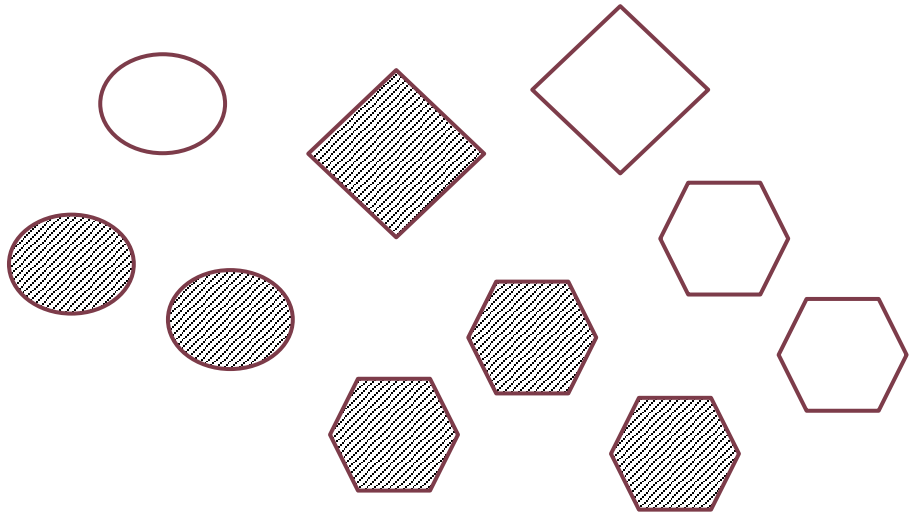
$$P(\text{oval or shaded}) = P(\text{oval}) + P(\text{shaded}) - P(\text{shaded oval})$$

Graphical example continued

- The example is a space of "weighted events":



- The equivalent example with "unweighted events" is:



- In this case, each event has the same probability, $P = 1/N$
- Often easier to use to help your intuition when manipulating probabilities

Combining probability: Joint probability

- Consider the model with two pushbutton boxes, and a set of trials where we push both buttons simultaneously
  - Label the boxes, and describe their behaviour by the random variables, *A* and *B*
  - Let P(A) mean P(*A*=1), likewise for P(B), and P(AB) mean that both P(*A*=1 and *B*=1)
  - IF the boxes are independent of one another (nothing connects them) then

    $$P(AB) = P(A)\,P(B)$$

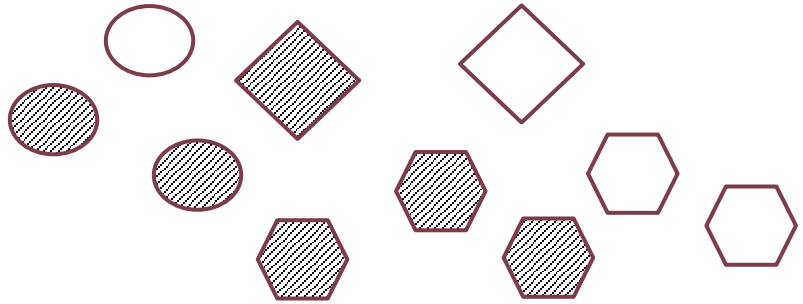
  - otherwise:

    $$P(AB) = P(A)\,P(B=1 \text{ given } A=1) = P(A)\,P(B|A)$$

  - the boxes are independent iff P(B) = P(B|A)

Graphical example of joint probability

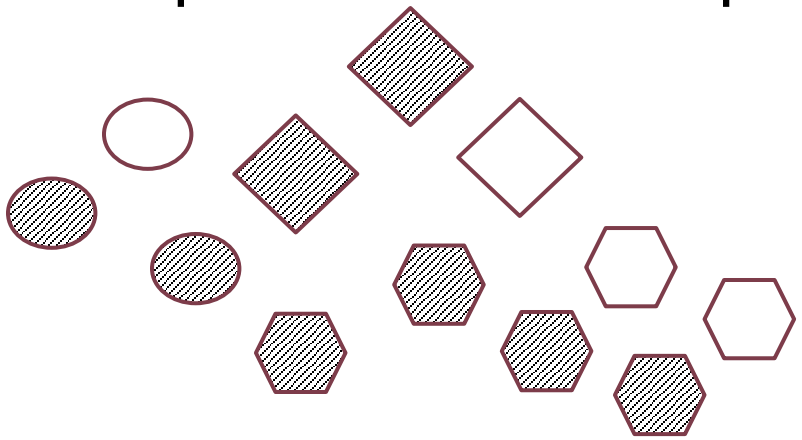
- In this space: $P(\text{shaded}) = 0.6$ and $P(\text{oval}) = 0.3$



Of the shaded shapes, 2 of 6 are ovals, so $P(\text{oval}|\text{shaded}) = 1/3$

$$P(\text{shaded oval}) = P(\text{shaded})P(\text{oval}|\text{shaded}) = 0.6/3 = 0.2$$

- In this space, shade and shape are independent:



$$P(\text{shaded oval}) = P(\text{shaded})P(\text{oval})$$

Bayes rule

- A simple formula, but very powerful

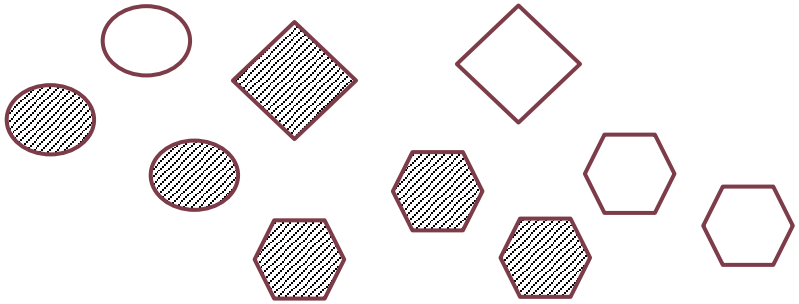
$$P(AB) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

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

- Applies to both Bayesian and Frequentist probability
- Plays a crucial role in defining how to update Bayesian probability after data is observed

Graphical example of Bayes rule

- Suppose an event from the space below is known to be an oval. What is the probability that it is shaded?



$P = 2/3$ obviously!

- Check with Bayes rule:

$$P(\text{shaded}|\text{oval}) = \frac{P(\text{oval}|\text{shaded})P(\text{shaded})}{P(\text{oval})} = \frac{2/6 \cdot 6/10}{3/10} = \frac{2}{3}$$

  - often the answer is not obvious and Bayes rule is needed

Subjective probability

- The concept of randomness does not enter into the definition of subjective probability
  - The fact that a system behaves in an unpredictable way can be considered to be due to our incomplete understanding of the system
  - It is not necessary to refer to repetitions

- Subjective probability is defined by the following:
  - It is a representation of degree of belief by real numbers
  - It has qualitative correspondence with common sense
    - Example: symmetric outcomes have equal probability
  - It takes into account all prior knowledge
  - It is mathematically consistent

Subjective probability example

- Dr. Jane performs an ALS test on John
  - let P(A) be the probability that John has ALS
  - let P(B) be the probability for John to have a positive ALS test result
- Dr. Jane's state of knowledge is the following
  - Prior to taking the test, she assigns the probability that John has ALS to be that of the general population

$$P(A)=0.001$$

  - The test is not perfectly predictive. From previous studies, Dr. Jane believes that

$$P(B|A) = 0.98 \text{ and } P(B|\bar{A}) = 0.03$$

  - The test comes back positive. Dr. Jane uses Bayes rule to deduce P(A|B)

Subjective probability example

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

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A})$$

$$= 0.98 \times 0.001 + 0.03 \times 0.999 = 0.03095$$

$$P(A|B) = \frac{0.98 \times 0.001}{0.03095} = 0.032$$

- Dr. Jane confidently tells John to not be concerned about the test result
- John may have a different prior belief and may have reason to be concerned…

Subjective probability

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

- P(A) is the prior belief
  - must always have a prior belief – subjective
- P(B|A) is the likelihood
  - probability of observing the data seen, given A
- P(B) is a normalizing factor
- P(A|B) is the posterior belief
  - indicates how the state of knowledge is updated as a result of observed data

Example shown graphically

- Prior to the test, the space is:

non-ALS

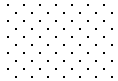
$P = 0.999$

ALS

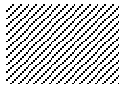
$P = 0.001$

- After the test, the space is:

- test

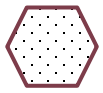


+ test



$P = 0.999 \cdot 0.97$
$= 0.96903$

$P = 0.999 \cdot 0.03$
$= 0.02997$



$P = 0.001 \cdot 0.02$
$= 0.00002$

$P = 0.001 \cdot 0.98$
$= 0.00098$

- Given positive test: $P(\text{ALS}) = 0.00098/(0.00098 + 0.02997) = 0.032$

Example from a frequentist viewpoint

- Bayes theorem can be applied in frequentist methods, provided all probabilities have a frequency interpretation (as is the case in this example)
  - P(A|B) is the probability that an individual drawn at random from a model population that would have a positive test result, actually has ALS

- In absence of prior knowledge P(A):
  - Bayesian approach cannot be used
  - Frequentist approach would result in the statement…

    "We reject the hypothesis that John does not have ALS at the 97% confidence level"
    - It sounds like the opposite of the Bayesian result!

Discussion 1: Lets Make a Deal!

- In a game show in the 1970's, a contestant is given the opportunity to pick one of three doors, behind one of which is the grand prize.
  - After the contestant picks a door, one of the remaining doors is opened (one without the prize). The contestant is then offered the chance to switch to the other door.
  - In order to maximize the chance of winning the grand prize, is it best for the contestant to
    - switch?
    - stay with the original choice?
    - it is the same either way?

Discussion 2: Money in envelopes

- Scenario 1:
  - Suppose you are told you can have a $10, or choose one of two envelopes (one has $5 and one has $20 inside).
    - What is the best strategy?
- Scenario 2a:
  - Suppose two envelopes are prepared by Jane who says that one has twice as much money as the other. You select one envelope but before opening it you are given the option to switch to the other envelope.
    - What is the best strategy?
- Scenario 2b:
  - Suppose envelopes are prepared as in 2a, but you are allowed to open the first envelope before deciding to switch, and find $10 in it.
    - Compare this to scenarios 1 and 2. Is there an optimal strategy?

Discussion 3: Survey

- Suppose that 1% of Canadians have a degree in physics and 80% of these physicists were men. A person is called at random by a polling agency who wants to find out about people's University degrees.
  - What is the probability that the person called has a physics degree?
  - If the person called is a man, what is the probability that he has a physics degree?
  - And if she is a woman?
  - Are "male" and "physics degree" independent?
  - Might they be independent in other populations?

Discussion 4: Weather forecasting

- Forecasts for the next day's weather includes the probability of precipitation (p.o.p.)
  - If one meteorologist states that the p.o.p. is 70% and another states that the p.o.p. is 40%, is at least one of the two necessarily wrong?
  - If you were asked to judge which of the two meteorologists had the better understanding of weather patterns by only looking at their past estimates of p.o.p. and the actual observations of precipitation, how would you decide?

Question: Two Dice

- In a simple game with dice, your score is the sum of the two dice points. If the two dice show the same number, your score is doubled.
  - Suppose after throwing the dice, you don't see what happened, but you are told that your score is 8.
- What is the probability that you rolled a "double"?

Question: French language in US/Canada

- In Canada there are about 31 million people, and according to the 2011 census, 22% speak French at home. In the US there are about 314 million people and 0.67% of them speak French at home.

- A person is selected from US/Canada "at random" and is found to speak French at home.
  - What is the probability that the person is from Canada?

Describing data and distributions

D. Karlen / University of Victoria and TRIUMF

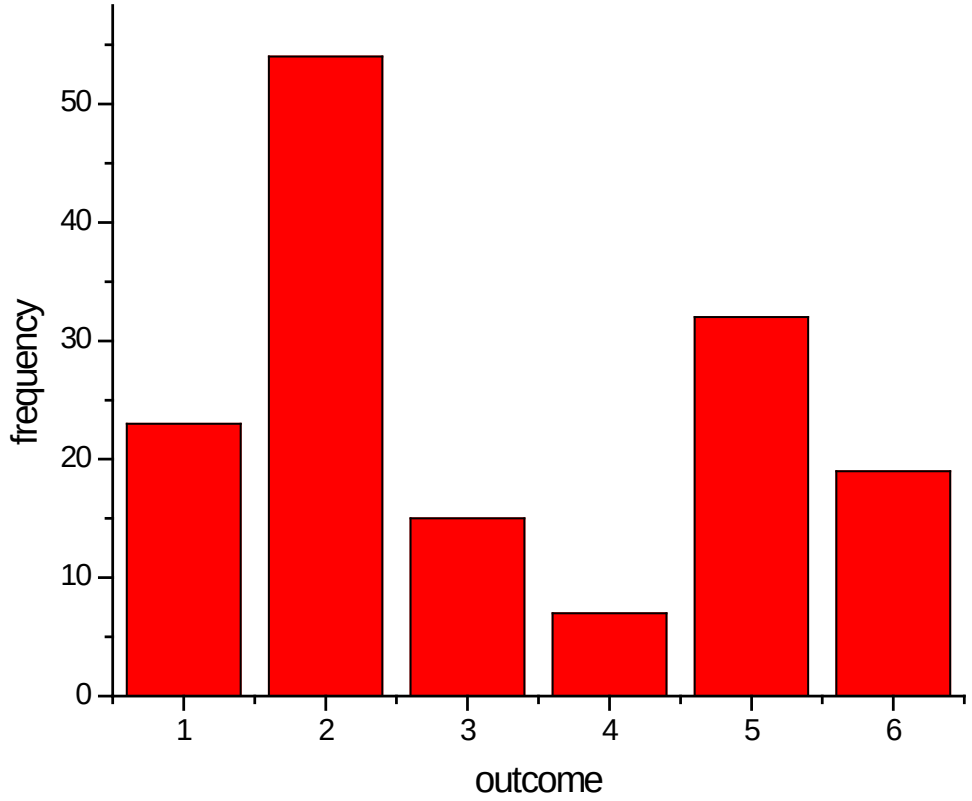
Describing data

- In the physical sciences, we learn by performing experiments
  - To learn something new and significant we often need to work at the limits of sensitivity → the outcomes of experiments are affected by factors outside of our control
    - our models include a random component to account for these factors
    - we repeat measurements to better understand these factors and to reduce their effect on our experimental outcomes

- Two approaches are used to give general impressions of repeated outcomes of an experiment:
  - Make a graphical representation of the data
  - Calculate some "descriptive statistics"

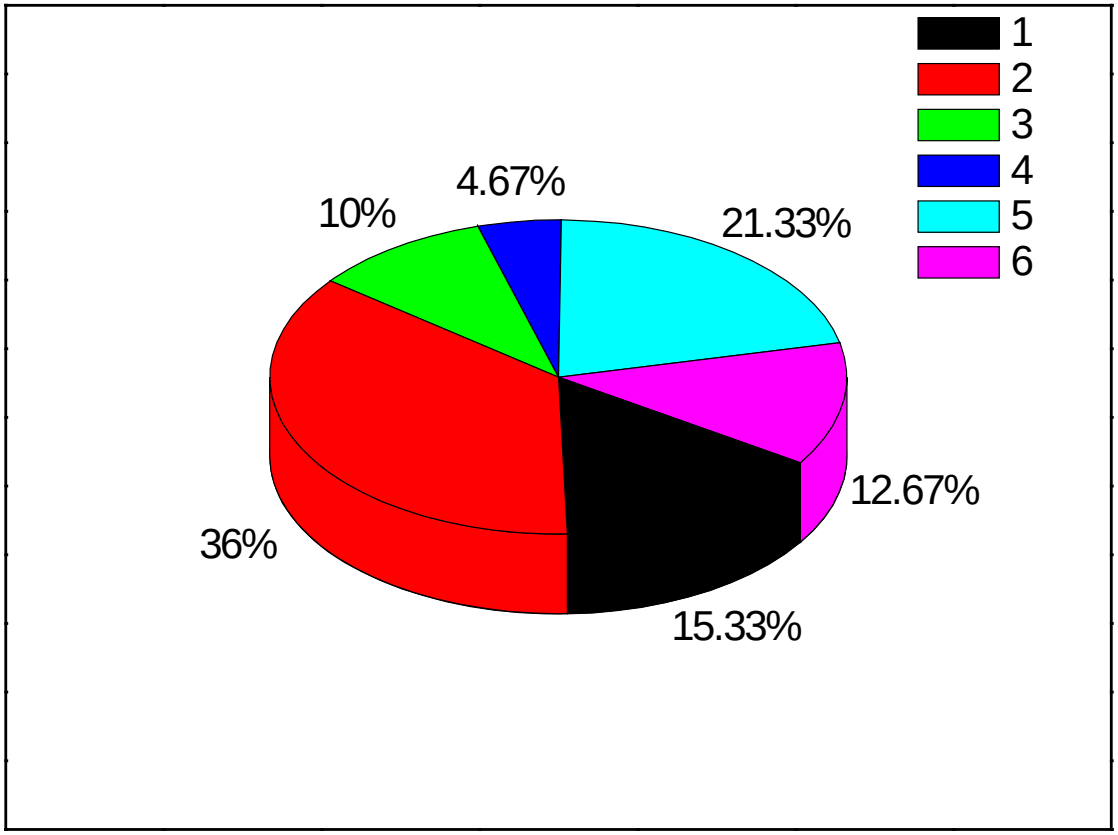
Histograms

- A graph that indicates the frequency of observing different outcomes from your experiment
  - it shows the "distribution" of the outcomes



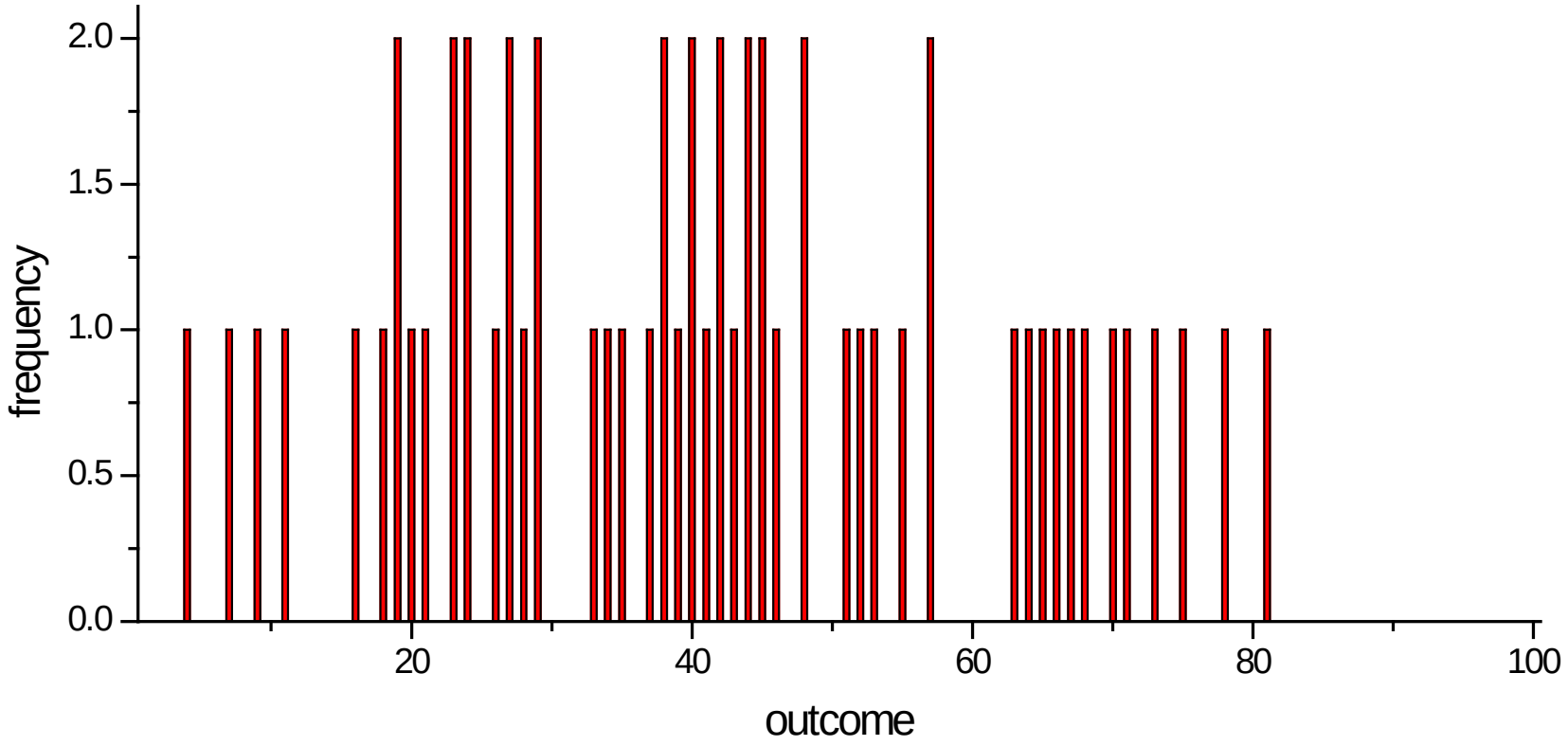
Alternatives to histograms

- Pie charts express the same information
  - Warning – you may loose scientific credibility if you use these!



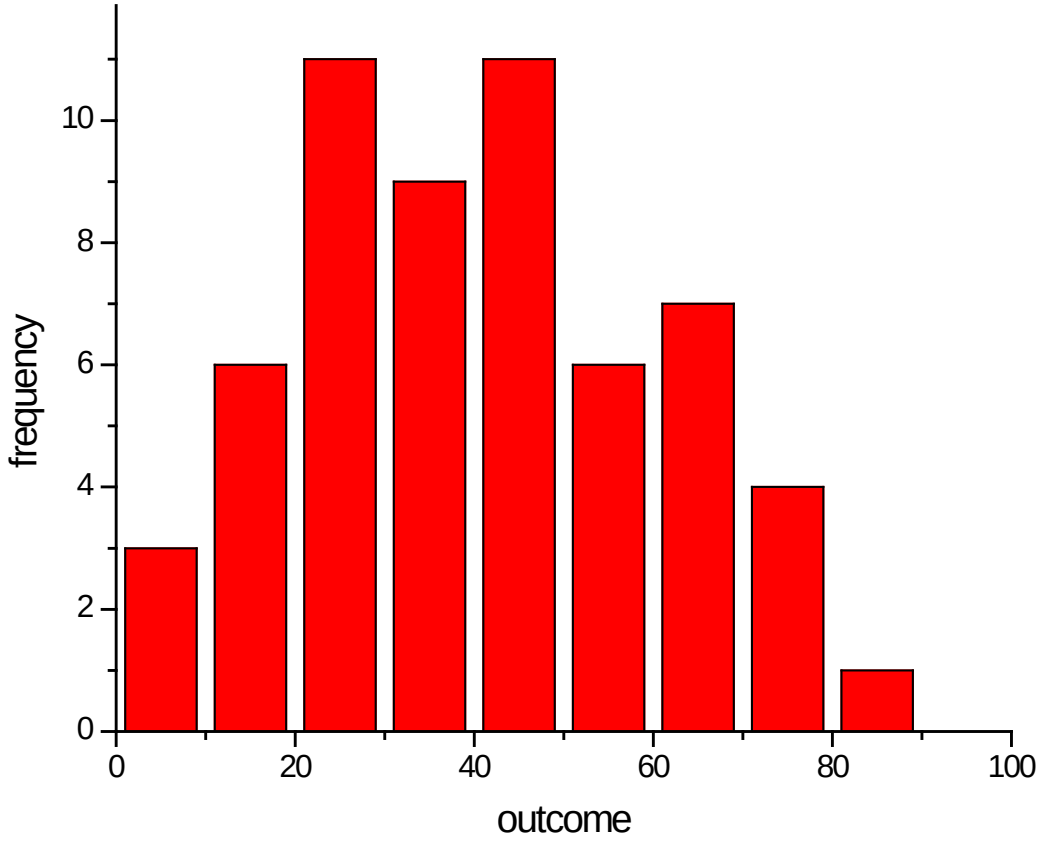
Histograms

- If your data has a very large number of possible outcomes, you may need to combine them into "bins"
  - So that at least some bins have many entries



Histograms

- If your data has a very large number of possible outcomes, you may need to combine them into "bins"
  - So that at least some bins have many entries



Descriptive statistics

- Sample mean (aka average)
  - Most meaningful single number to describe the outcomes of your experiment

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

- Sample variance
  - A number that indicates the spread of the outcomes

$$V_x = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \overline{x^2} - \bar{x}^2$$

  - The sample standard deviation is $\sigma_x = \sqrt{V_x}$

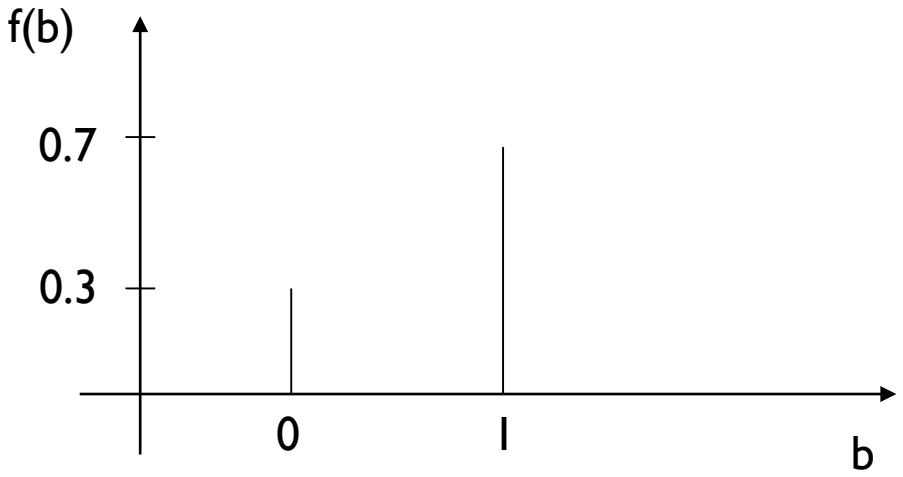
Relating data to models

- Due to uncontrollable factors repeated observations yield a distribution of results
- To model this behaviour, you can include a random component
  - Any particular outcome has a certain probability of occurring
- Suppose our model has the following probabilities: P(*B*=0) = 0.3 and P(*B*=1) = 0.7.
  - A convenient way to express this is to say the *B* has the following probability distribution:

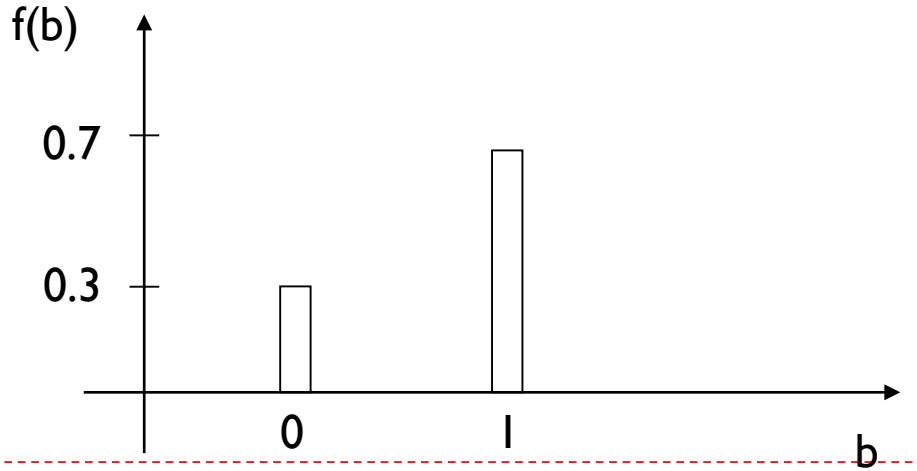
$$P(B = b) = f(b) \qquad f(b) = \begin{cases} 0.3 & b = 0 \\ 0.7 & b = 1 \end{cases}$$

Probability distributions

- The distribution can be drawn as follows:



- or :



Continuous random variables

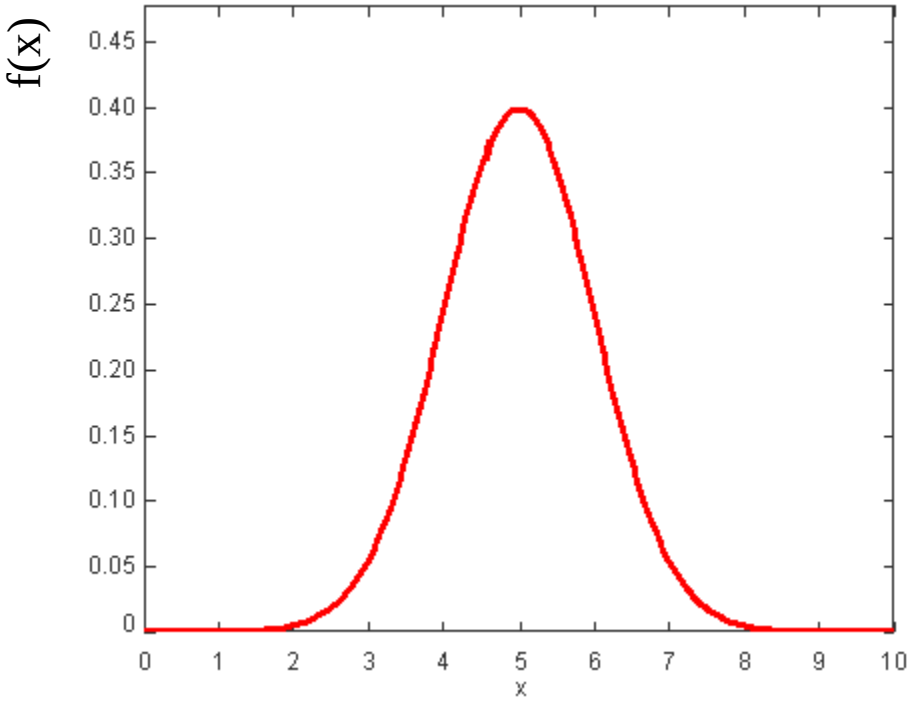
- Most experiments record quantities that appear to be continuous; for example, mass or temperature.
  - In reality, systems cannot produce all possible values, the values are quantized according to its precision.
  - Even so, for convenience, such systems are usually modeled by a continuous random variable; call it *X*.
  - A probability density function $f(x)$ is used express the behaviour of such a model:

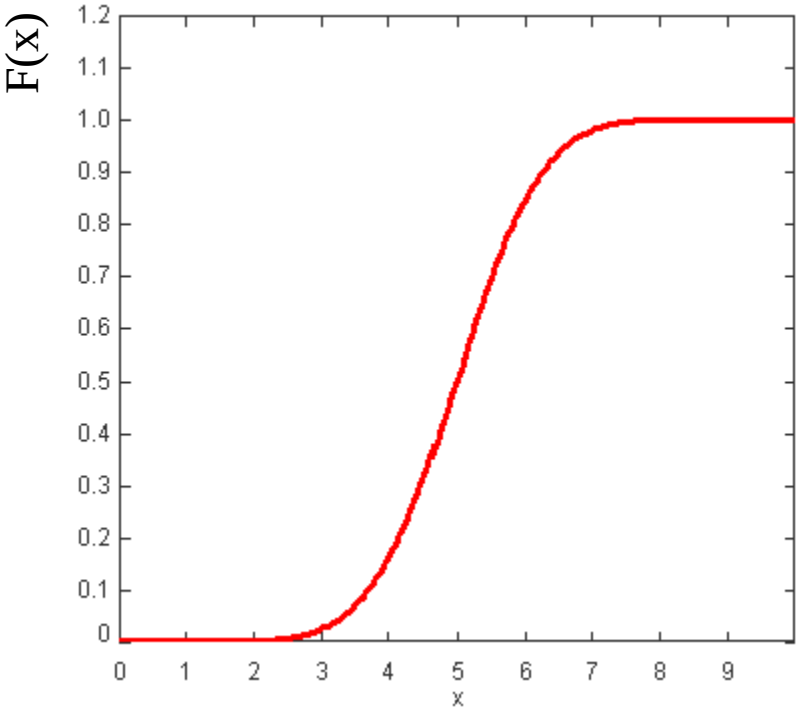
$$P(x < X < x + dx) = f(x)\,dx$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad F(x) = \int_{-\infty}^{x} f(x')dx' \qquad \text{cumulative distribution}$$

Probability density function

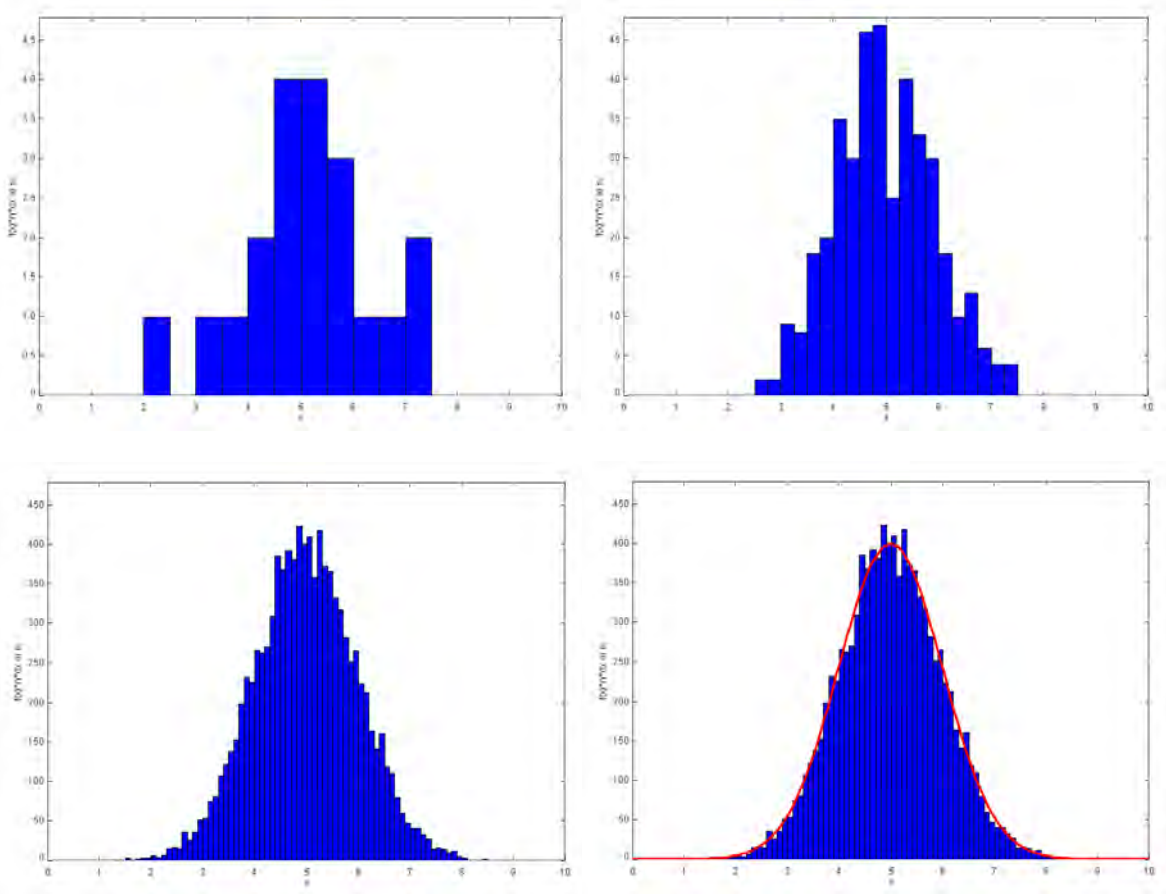
- Example: Normal pdf and its cumulative





Relation to histograms

- As more outcomes are collected, and the bin widths are reduced, the histogram takes the shape of the pdf



- The red curve in the final figure is the pdf multiplied by $n\Delta x$:
  - n =10,000
  - $\Delta x = 0.1$

Question: histograms and pdfs

- When overlaying a histogram with a pdf, you must multiply the pdf by $n\Delta x$, where n is the number of entries in the histogram and Δx is the bin width.
  - Explain why this is the correct "scaling factor"

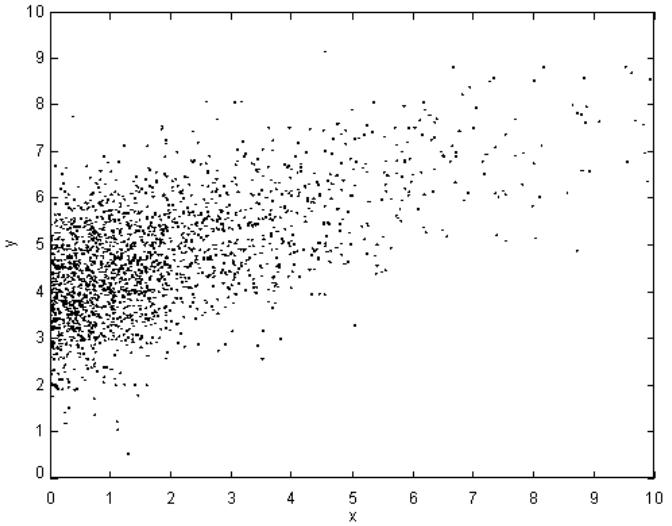
- Explain what the procedure is to overlay data with a cumulative distribution, for comparison

Question: example pdf

- Suppose X is a random variable whose probability density is zero at $x \leq 0$ and rises linearly with x, except for $x > 10$ where the density is zero again.
  - What is the mathematical form for the pdf, $f(x)$?
  - What is probability for X to be in the range $(4,6)$?
  - What is the probability for $X = 5$?

Data with more than one observable

- Some experiments have more than one observable
  - Outcomes of repeated experiments can be shown as a scatter plot:

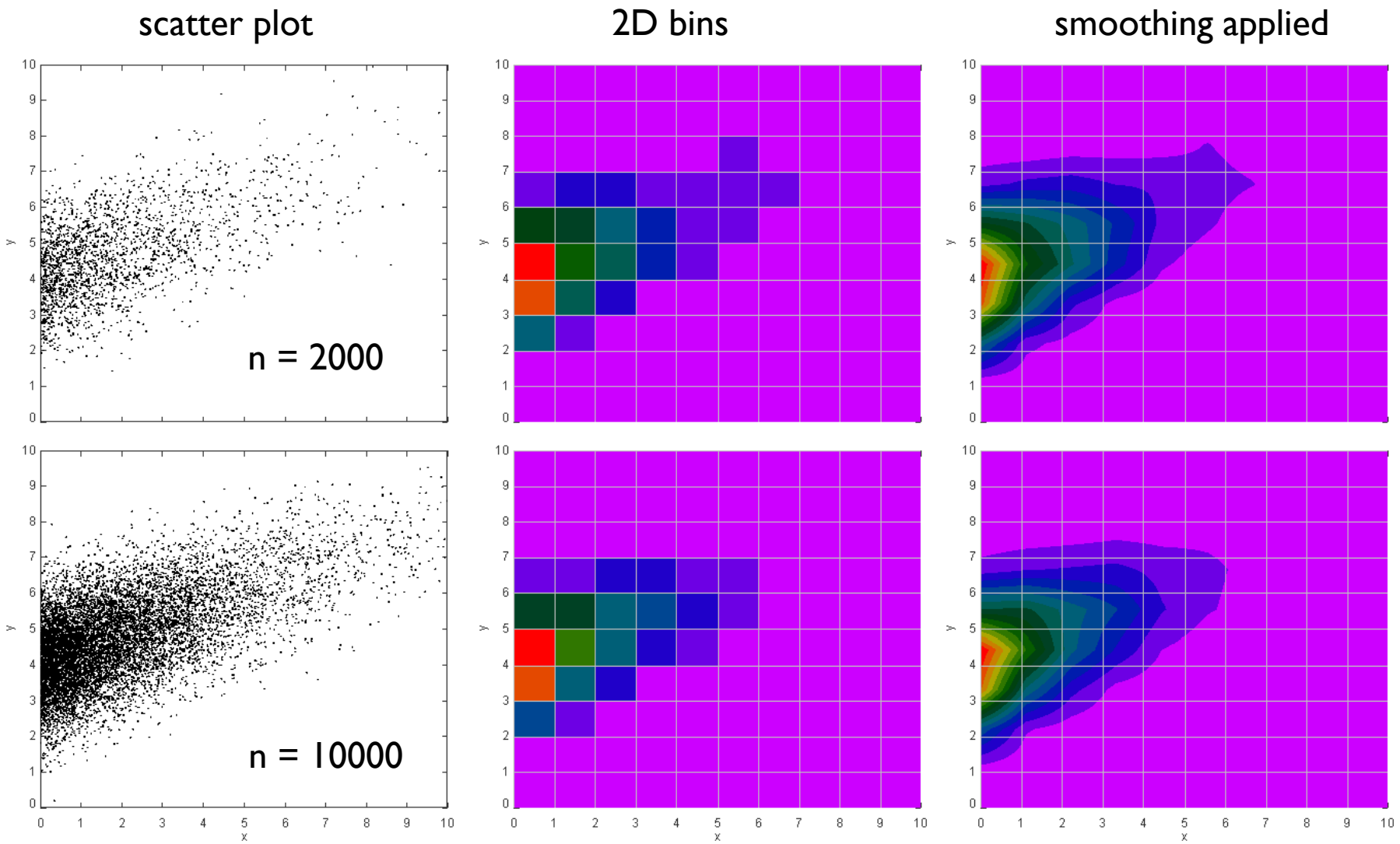


- The average value of x or y, or their sample variances are useful descriptive statistics
- The sample covariance V_{xy} and correlation coefficient, ρ, describe the "shared variation":

$$V_{xy} = \mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}$$

$$\rho = V_{xy} / \sqrt{V_x V_y} = V_{xy} / (\sigma_x \sigma_y)$$

Presenting multivariate data in 2D bins



Modeling with multivariate pdfs

- With two observables, the experiment is modeled by pairs of random variables X and Y

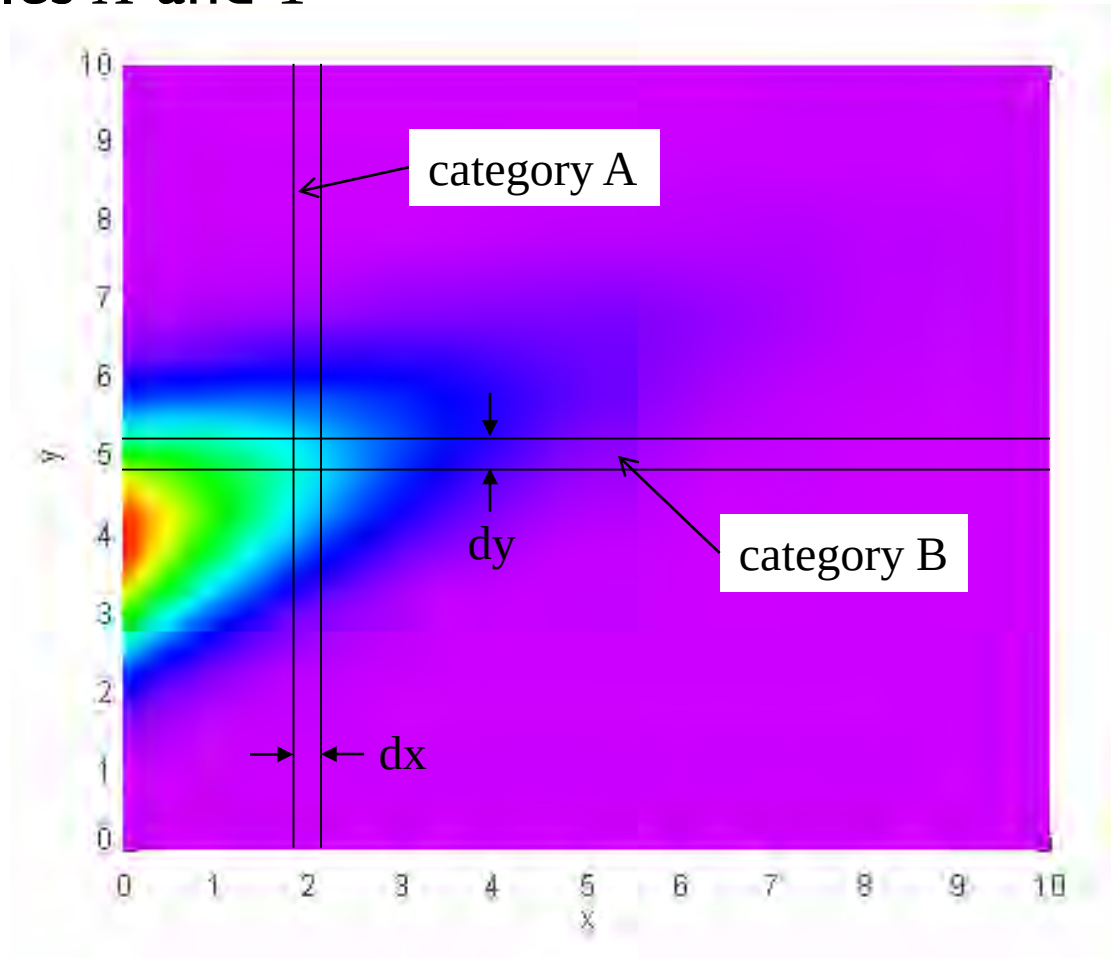
ejs_plot2D.jar

$$P(AB) = f(x,y)dxdy$$

joint pdf

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

  - probability density shown by a colour scale in this example

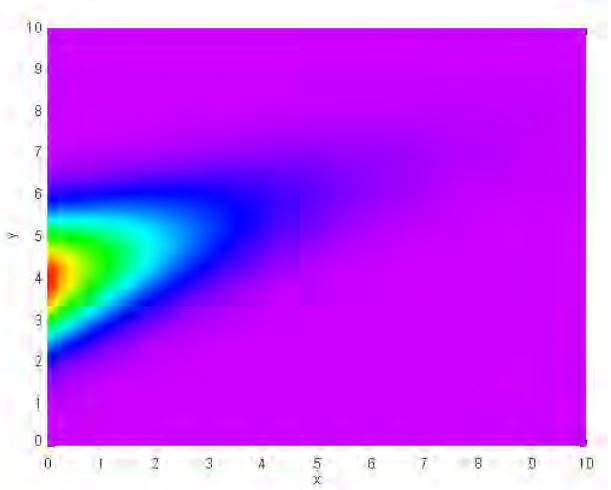


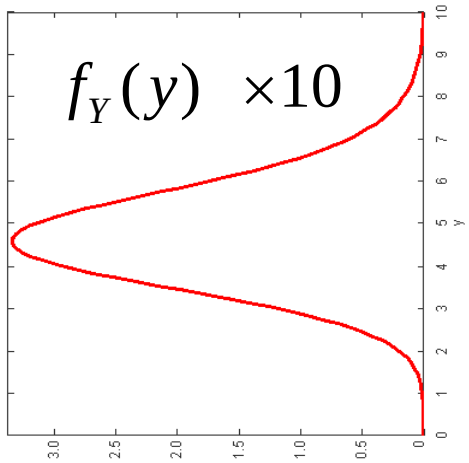
Marginal pdfs

- Projections onto x and y axes:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

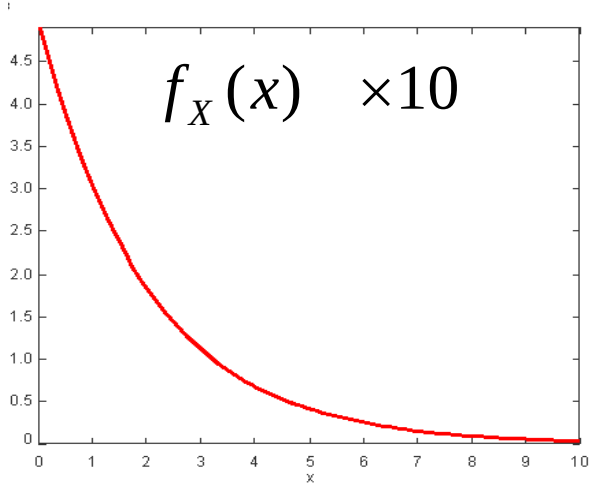
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$





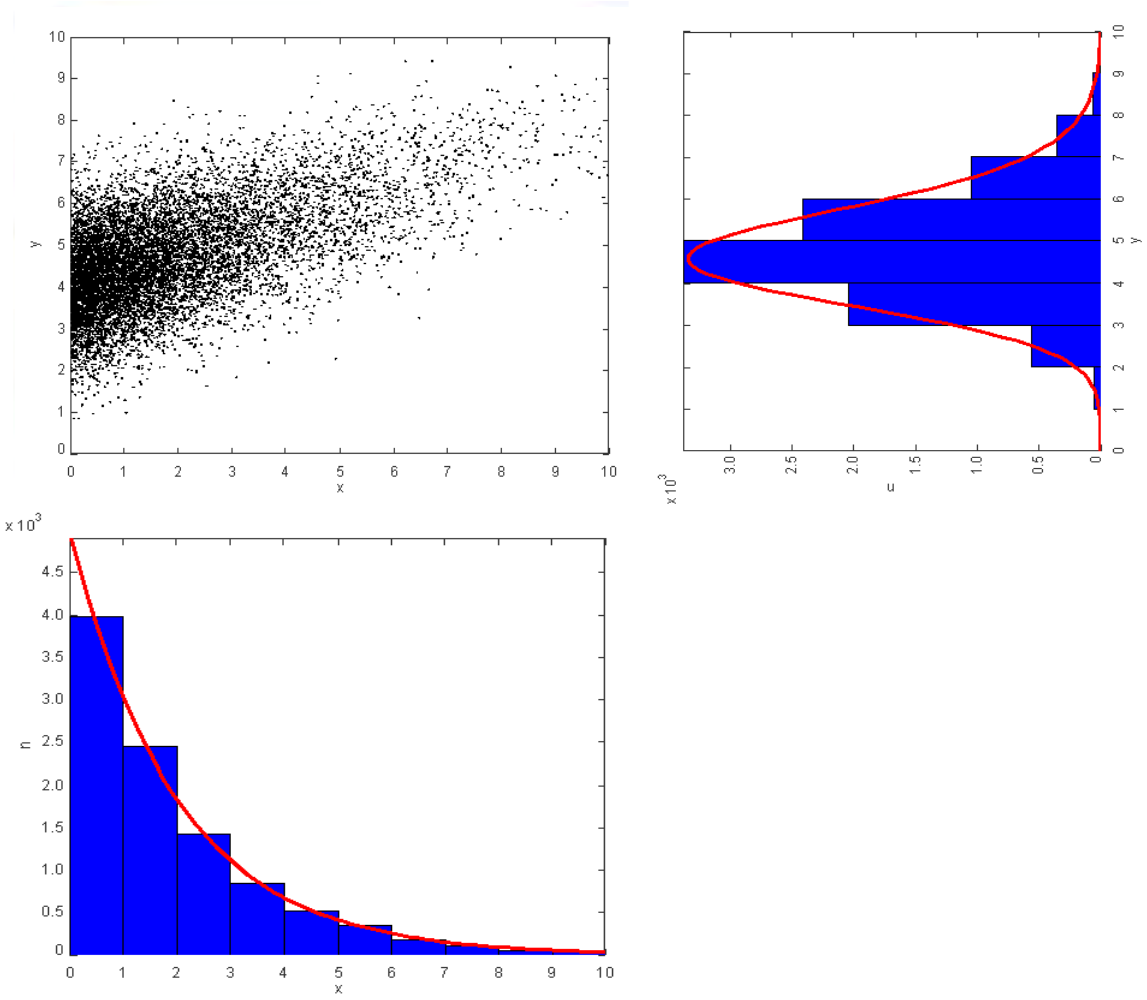
- Recall: X and Y are independent iff

$$f(x,y) = f_X(x)\,f_Y(y)$$



Marginal distributions

- Project data onto *x* and *y* axes

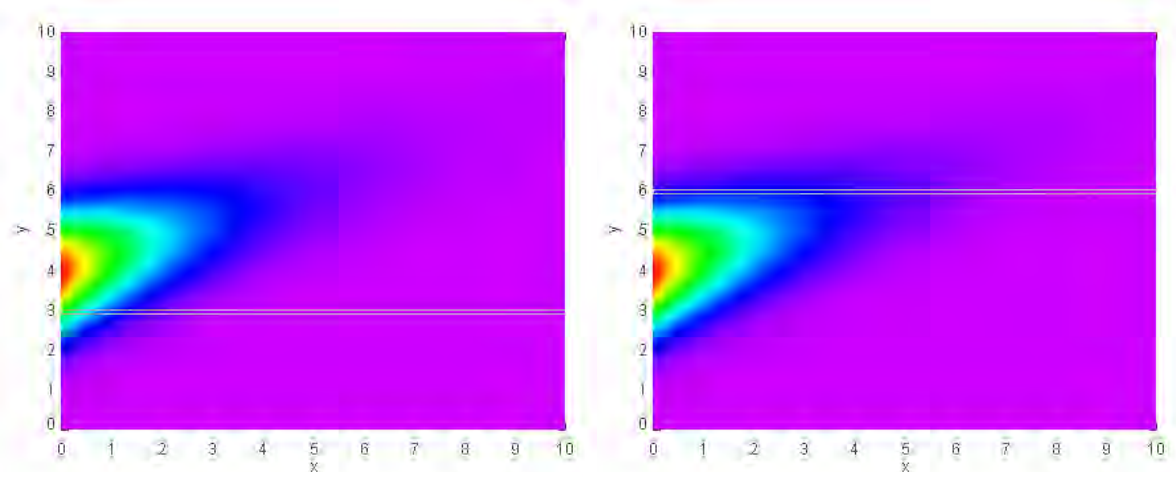


Conditional pdfs and distributions

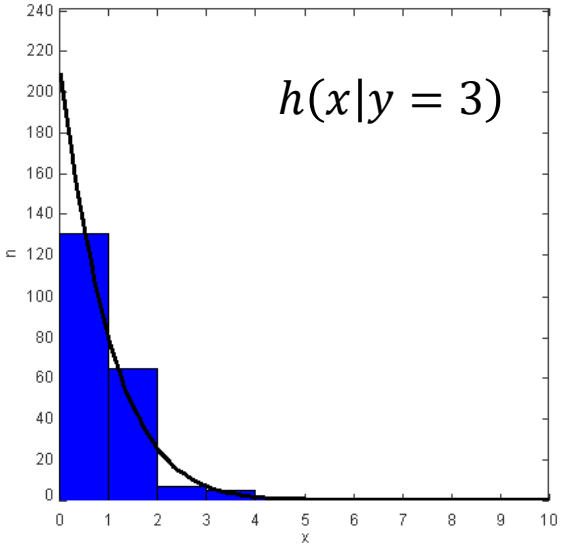
- Conditional pdf: a normalized slice of the joint pdf:

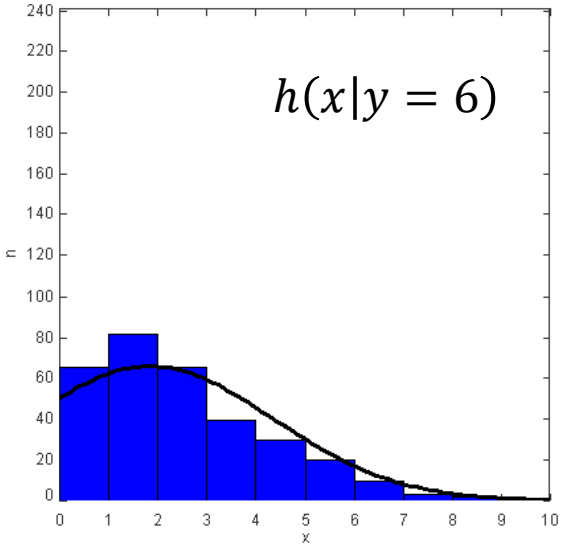
$$h(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

$$\int_{-\infty}^{\infty} h(x \mid y)\, dx = 1$$



- Conditional distribution:
  - distribution of one variable for data that satisfies conditions on other variable





Question: Independence

- If X and Y are independent, what can be said about the conditional pdfs, $h(x|y)$ and $h(y|x)$?

- Use the "Plot2D" app and adjust the pdf parameters
  - For each pdf type, what parameter values yield random variables X and Y that are independent?

Question: Units

- Suppose the random variables, M corresponds to mass in kg, and T time in seconds. What are the units of the following pdfs?
  - $f(m)$
  - $g(m,t)$
  - $h(m|t)$
  - $F(m)$ (the cumulative distribution of $f(m)$)
  - $g_M(m)$ (the marginal pdf of g for m)

- What possible values can the following take?
  - $f(m)$
  - $g_M(m)$

pdf properties: Expectation value

- Given a continuous random variable *X* described by the probability density, *f*(*x*):

$$P(x < X < x + dx) = f(x)\,dx$$

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

the expectation value of *X* is defined as

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

  - also known as the mean of *X*
  - often denoted by μ

pdf properties: Variance

- The variance of *X is defined as*:

$$V_X = E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty}(x-\mu_X)^2 \; f(x)\,dx$$

$$= E\left[X^2\right] - \mu_X^2$$

- The "descriptive statistics" of observed data, introduced at the beginning of this section, are estimates of properties of the pdf that produced the data:

descriptive statistic	pdf property it estimates
average	expectation value
variance	variance

pdf properties: Covariance

- The covariance of two random variables X and Y is defined to be:

$$V_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\,f(x,y)\,dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\; f(x,y)\,dxdy - \mu_X\mu_Y = E\left[XY\right] - \mu_X\mu_Y = \mathrm{cov}(X,Y)$$

- The covariances are often collected into a covariance matrix (aka the error matrix)

$$V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{XY} & V_{YY} \end{pmatrix}$$

pdf properties: Correlation coefficient

- The covariance is the variance that is linearly shared between two random variables
- The correlation coefficient is defined by:

$$\rho_{XY} = \frac{V_{XY}}{\sigma_X \sigma_Y}$$

a dimensionless measure of correlation, and lies between -1 and 1.

- Examples:

$$X = Y \Rightarrow \rho_{XY} = 1$$
$$X = -Y \Rightarrow \rho_{XY} = -1$$

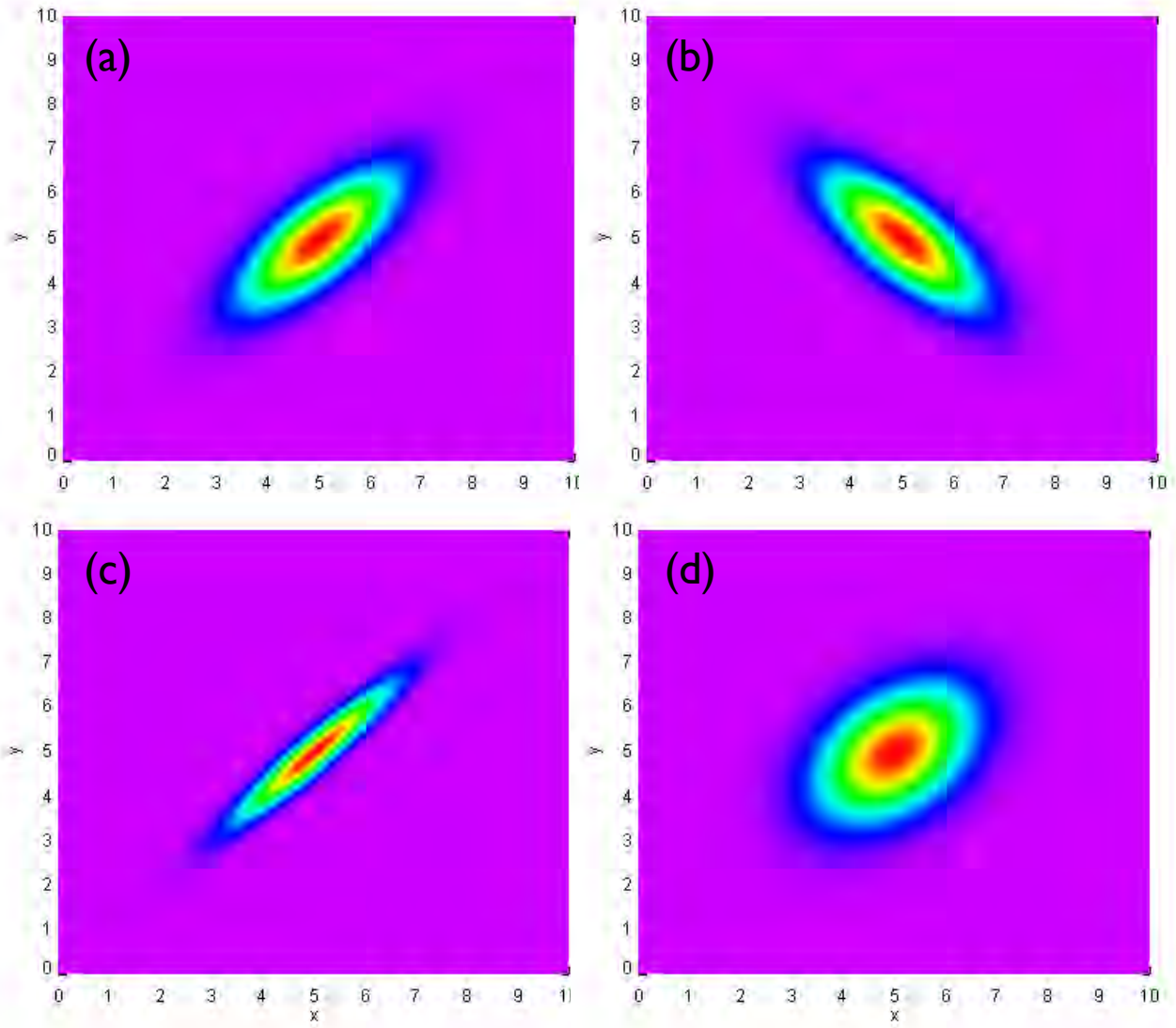
"completely (anti-)correlated"

$$E[XY] = E[X]E[Y] \Rightarrow \rho_{XY} = 0$$

"uncorrelated"

Correlation coefficient

- More examples:



Questions:
Which of the pdfs describe random variables that:

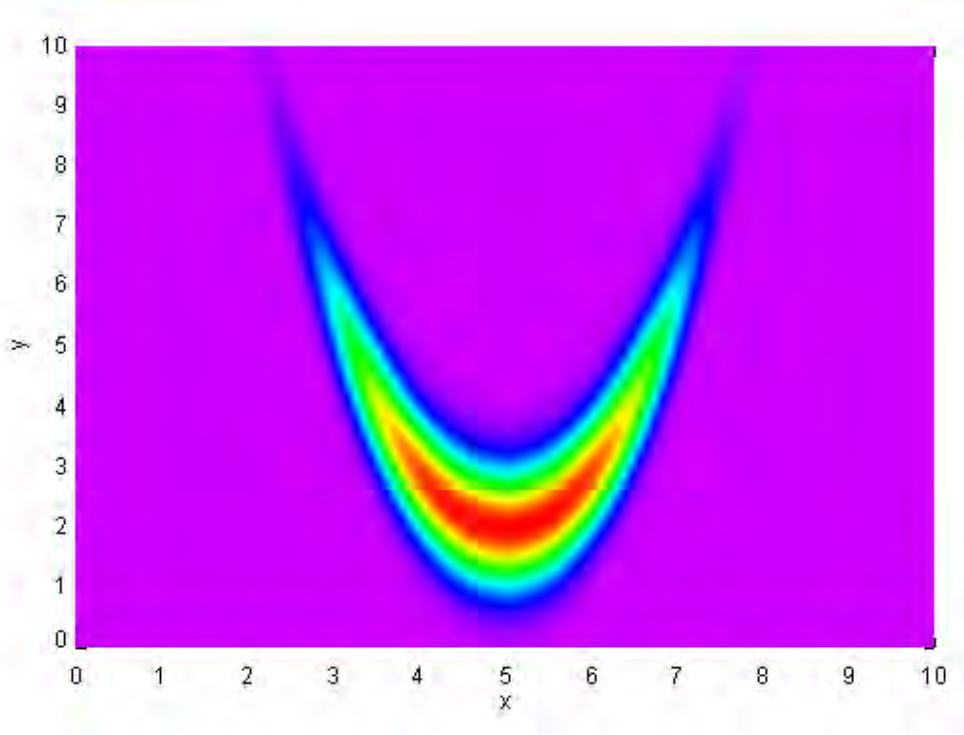
- are correlated?
- are positively correlated?
- which one has the strongest correlation?

Correlation coefficient

- Common misunderstanding:
  - If X and Y are independent, then $\rho_{XY} = 0$
  - If $\rho_{XY} = 0$, then X and Y are independent?
    - not necessarily!

Example:

$E[XY] = E[X]E[Y]$



  - do not confuse the two terms: uncorrelated and independent have different meanings

Question: example pdf

- Suppose X is a random variable whose probability density is zero at $x \leq 0$ and rises linearly with x, except for $x > 9$ where the density is zero again.
  - What is expectation value and variance of X ?
  - What is the probability for X to exceed the expectation value?
  - What is the probability for X to be within one standard deviation of the expectation value?

Question: correlation of mass and density

- Consider an experiment that measures the mass and density of people, by weighing them and measuring the displacement of water when they are submerged in a pool. Would the correlation be positive or negative in the models that describe this experiment for the following two cases?
  - A: repeated pairs of measurements of the same individual, done over a short period of time (during which time the person does not eat, etc.)
  - B: single pairs of measurements for a group of individuals

Question: covariance of measurements

- Consider measurements of the resistance of a resistor undertaken by two people, one right after another.
  - Suppose repeated measurements give different results because:
    - variable quality of the contact between ohm-meter and resistor
    - slow variations of the temperature of the ohm-meter
  - To model this situation:, use two random variables to represent future measurements by the two people. Assume that the two measurements are done at the same temperature.

$$X = \mu + C_X + T \qquad E[C_X] = 0 \qquad E[T] = \mu_T$$

$$Y = \mu + C_Y + T \qquad E[C_Y] = 0$$

  - What is the covariance, V_{XY} ? Answer: V_T

Question: covariance for different sensitivity

- Suppose two ohm-meters meters have different linear temperature dependencies. If the random variables, X and Y, represent hypothetical measurements from the two meters, their expectations at temperature t are:
  - $E[X] = \mu + a(t - t_0)$
  - $E[Y] = \mu + b(t - t_0)$
- In absence of temperature variations, the ohm-meters have independent variation due to unknown factors:
  - $V_X = V_Y = \sigma^2$
- In a situation where the temperature variation is:
  - $\mathrm{E}[T] = t_0 + 1$ and $V_T = \sigma_T^2$
- Calculate: $E[X]$, V_X, and V_{XY}

Functions of random variables

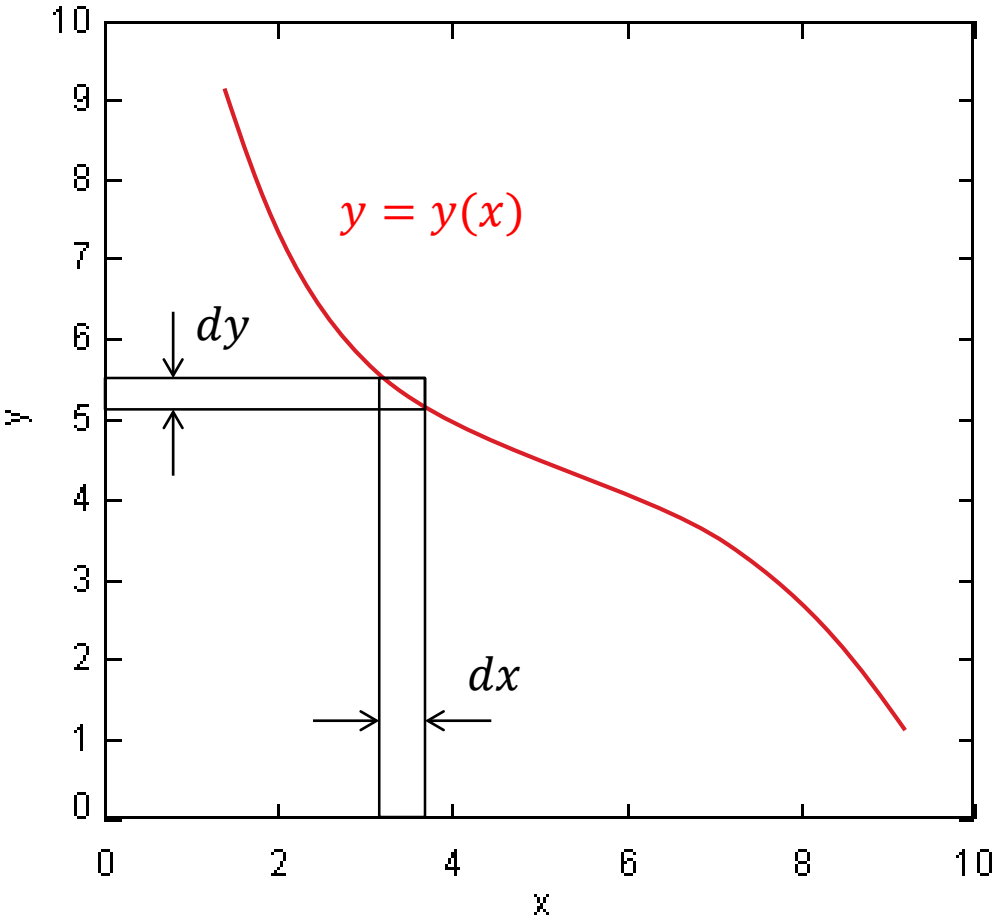
- Some systems can be described by outcomes of a function of a simple random variable
- A function of a random variable is also a random variable
  - Consider a continuous random variable X with pdf $f(x)$ and a random variable, Y, that is formed by some mathematical operation on X, ie. $Y = Y(X)$.
    - A simple example: $Y = a + bX$
  - What is the pdf of Y, $g(y)$?

  - Answer: Use the fact that the probability content of a range of outcomes are related:

$$P(x < X < x + dx) = f(x)\,dx$$
$$P(y < Y < y + dy) = g(y)\,dy$$

Functions of random variables

- If the function is monotonic (single valued inverse):
  - The probability for X to be within $[x, x + dx]$ is the same as the probability for Y to be within $[y, y + dy]$



$$g(y)\,dy = f(x)\,dx$$

$$g(y) = f(x)\left|\frac{dx}{dy}\right|$$

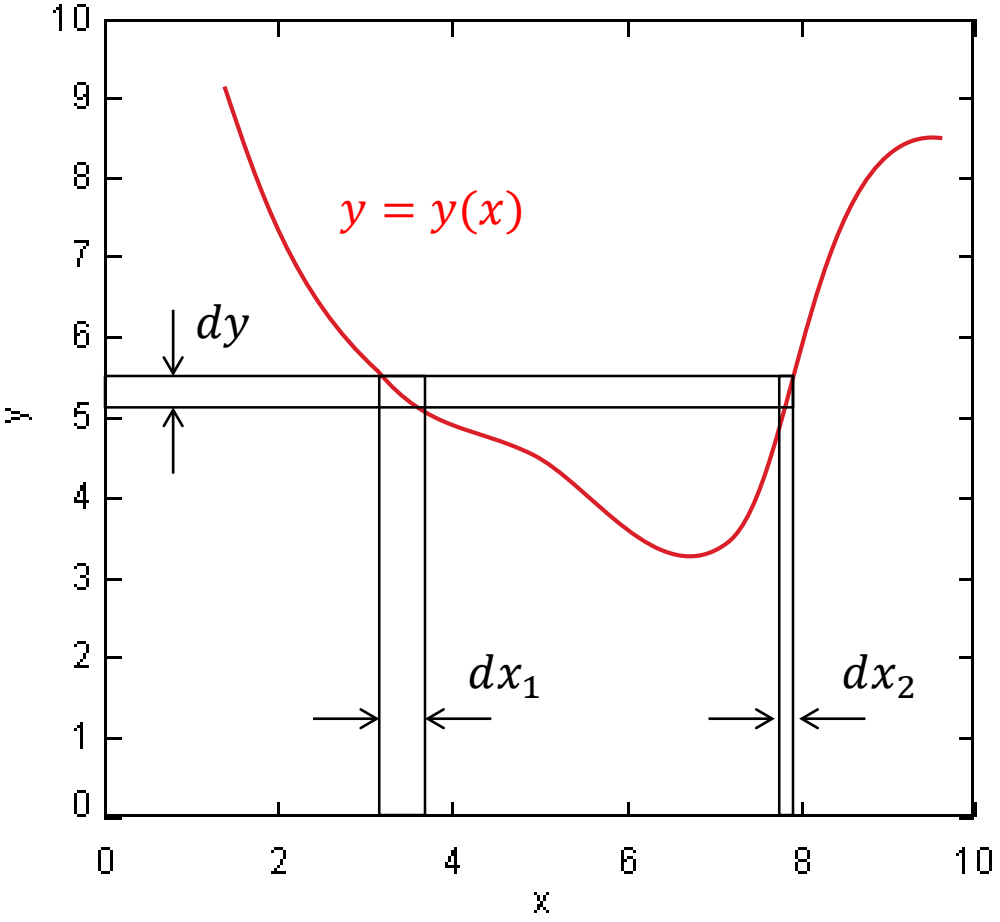
$$= f(x)\left|\frac{dy}{dx}\right|^{-1}$$

Functions of random variables

- If the function inverse is multivalued:
  - Include all solutions

$$g(y) = f(x_1)\left|\frac{dy}{dx}\right|_{x=x_1}^{-1}$$

$$+ f(x_2)\left|\frac{dy}{dx}\right|_{x=x_2}^{-1}$$



Convolution of random variables

- Convolution is the combination of more than one random variable to form a new random variable
  - The most common convolution is adding two random variables

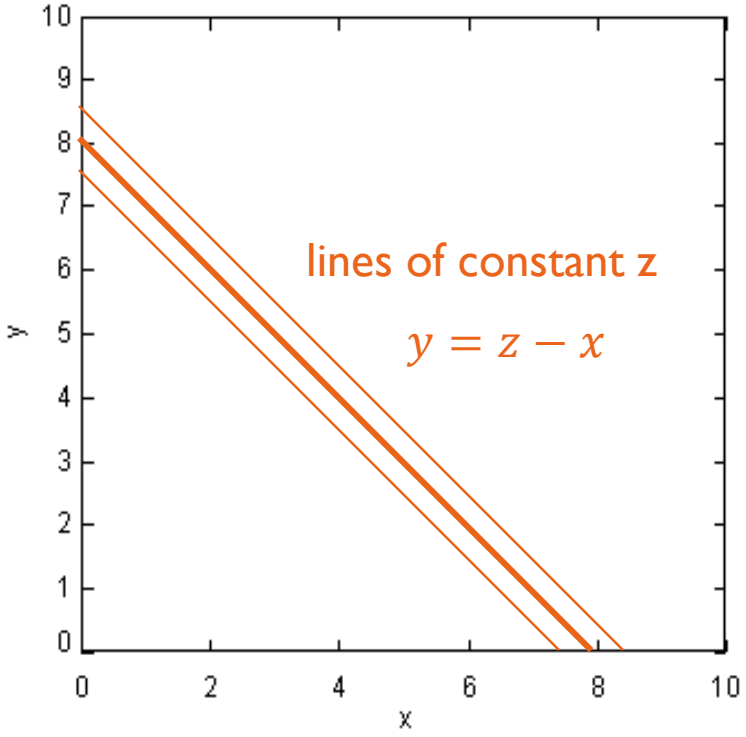
$$Z = X + Y$$

  - Consider the outcomes of X and Y that produce an outcome of Z in the range $[z, z + dz]$:

$$P(z < Z < z + dz) = h(z)\,dz$$
$$= \left(\int_{-\infty}^{\infty} f(x, z - x)\,dx \right) dz$$

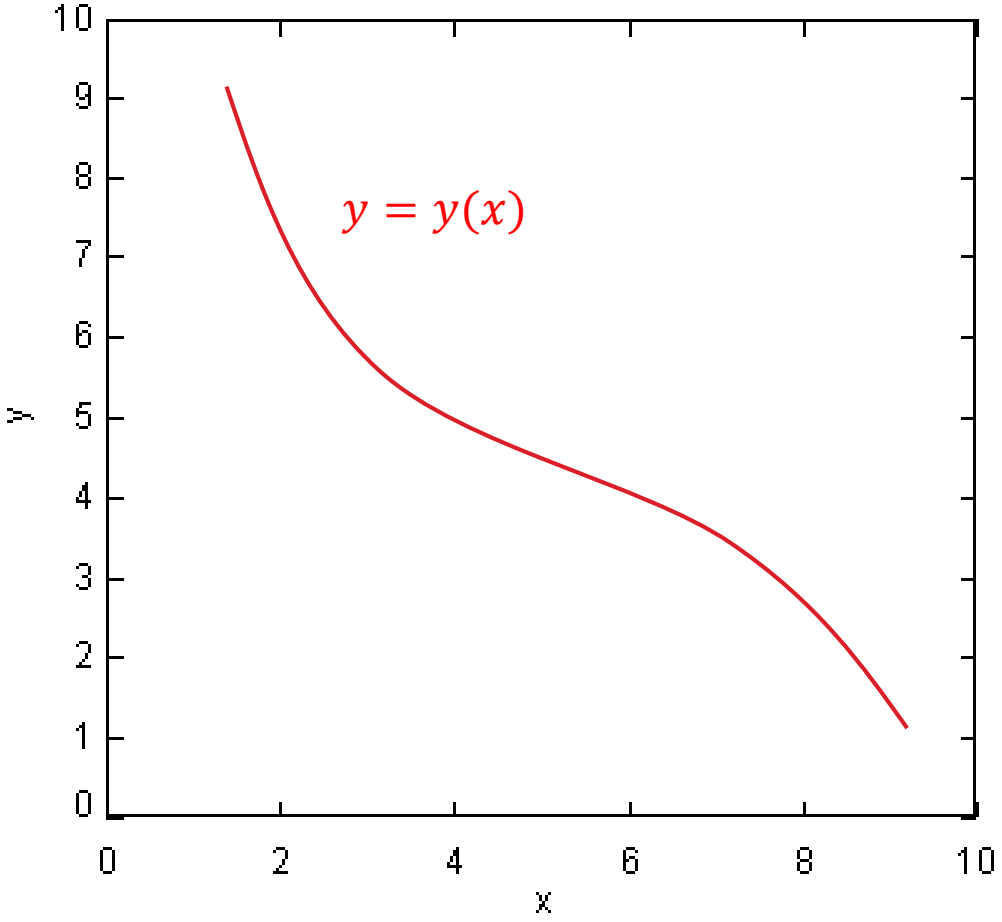
  - If X and Y are independent:

$$h(z) = \int_{-\infty}^{\infty} f(x) g(z - x)\,dx$$



Question: function of a random variable

- If X has uniform probability density in the range $[2,8]$ and zero density elsewhere, and Y is a function of X as illustrated here, describe the general features of the probability density, $g(y)$



Question: linear function of a RV

- Consider the simple function of a random variable X $Y = a + bX$, where a and b are constants.
  - What is the expectation values and variance of Y in terms of the corresponding properties of X?
  - Compare with a derivation directly from the definitions of expectation value and variance.

Question: quadratic function of a RV

- Consider the quadratic function of a random variable X $Y = a + bX^2$, where a and b are constants.
  - Suppose X has uniform density between $(0,10)$ and zero density elsewhere, what is the probability density for Y?
  - What is the expectation values and variance of Y?

Question: Convolution

- Suppose X and Y are independent random variables whose probability density are both constant in the range $(0,10)$ and zero elsewhere
- What is the probability density for $Z = X + Y$?

- Suppose the random variables are not at all independent, but instead, $X = Y$. What is the probability density for $Z = X + Y$ now?

ejs_probDist.jar

Special probability distributions

D. Karlen / University of Victoria and TRIUMF

Special probability distributions

- There are several probability distributions that are particularly useful when developing models of physical systems with random variables:
  - Binomial distribution
  - Multinomial distribution
  - Poisson distribution
  - Uniform distribution
  - Exponential distribution
  - Gaussian (normal) distribution
  - Chi-square distribution

Binomial distribution

- Used for modeling repeated independent observations (under identical conditions) of an non-predictive system that has two possible outcomes:
  - success or failure
  - life or death
  - heads or tails

- The model is defined by p (probability of success) and N (number of trials)
- Since each trial is independent, the result is given by the random variable n (number of "successes")
  - the order of successes is not relevant

Example:

- A simple model for 3 free throw attempts, each with probability 0.8 of success:

Results	Probability
0 0 0	0.2*0.2*0.2 = 0.008
0 0 1	0.2*0.2*0.8 = 0.032
0 1 0	0.2*0.8*0.2 = 0.032
0 1 1	0.2*0.8*0.8 = 0.128
1 0 0	0.8*0.2*0.2 = 0.032
1 0 1	0.8*0.2*0.8 = 0.128
1 1 0	0.8*0.8*0.2 = 0.128
1 1 1	0.8*0.8*0.8 = 0.512

n	Probability
0	1 * 0.008 = 0.008
1	3 * 0.032 = 0.096
2	3 * 0.128 = 0.384
3	1 * 0.512 = 0.512

Binomial distribution

- The binomial distribution is defined to be the probability of observing *n* successes in this model:

$$f(n \mid N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

  - since *n* is a discrete variable (integer), this is known as a probability mass function (instead of probability density function):

$$P(n = n) = f(n \mid N, p)$$

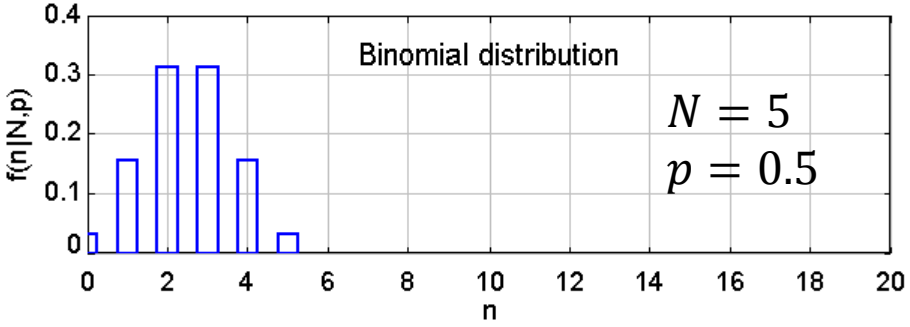
Properties of the binomial distribution

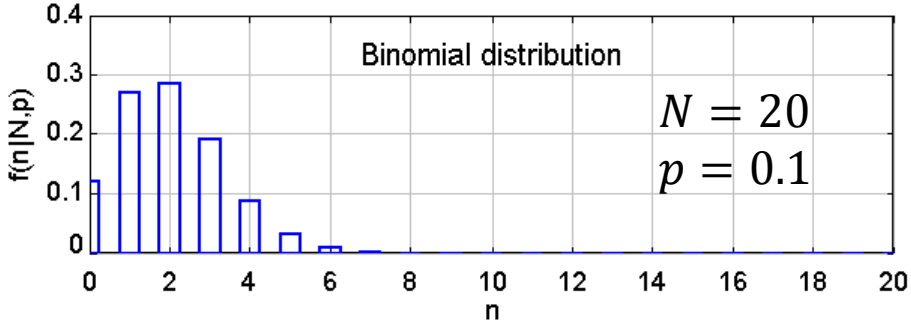
- Expectation value and variance:

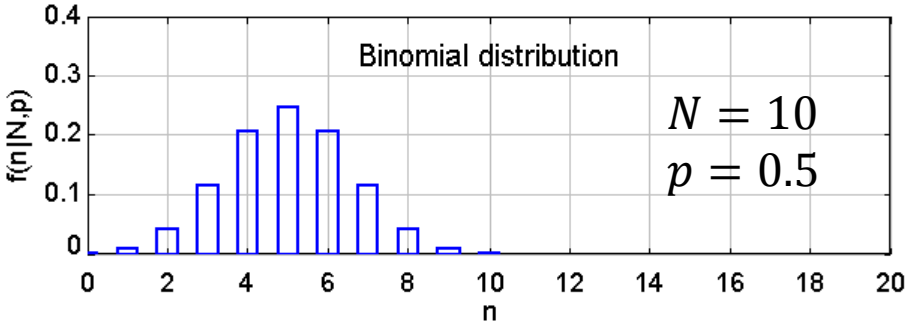
$$E[n] = \sum_{n=0}^{\infty} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} = Np$$

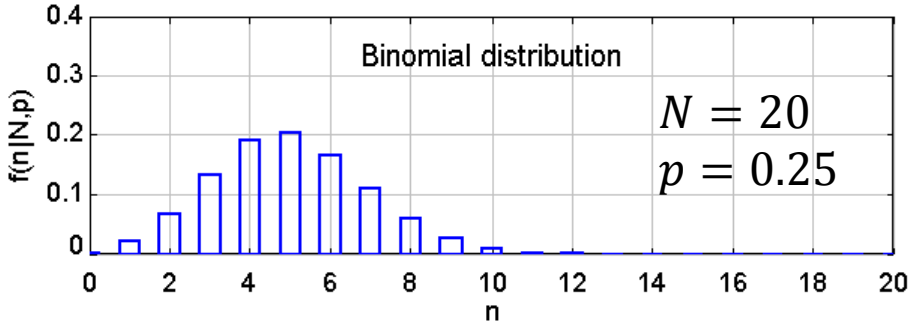
$$V[n] = Np(1-p)$$

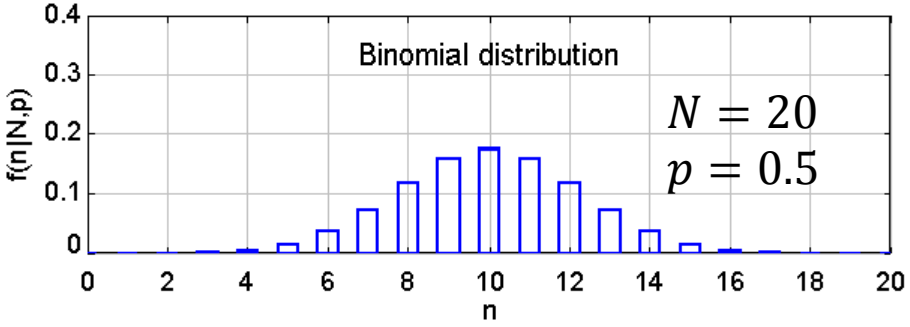
Examples of binomial distributions

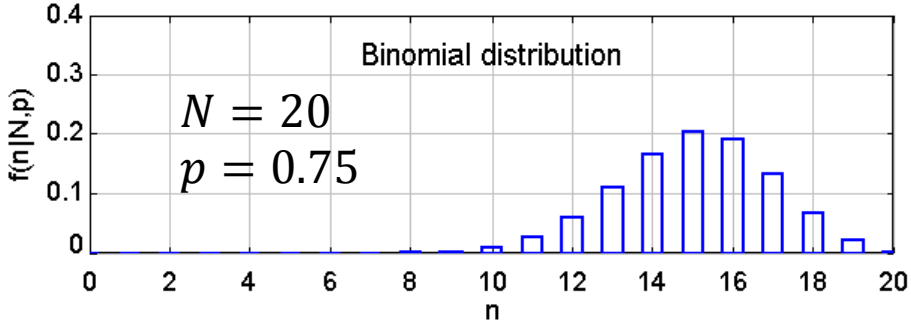












Multinomial distribution

- This is a generalization of the binomial distribution for systems with more than two possible outcomes.
  - If there are m possible outcomes,

$$f(n_1, n_2, \ldots, n_m | N, p_1, p_2, \ldots, p_m) = \frac{N!}{n_1!\, n_2! \ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}$$

  - Properties:

$$E[n_i] = Np_i$$

$$V_{ij} = \begin{cases} Np_i(1-p_i) & i = j \\ -Np_ip_j & i \neq j \end{cases}$$

Poisson distribution

- A limiting case of the binomial distribution
  - $N \to \infty, p \to 0, Np = \nu$

$$f(n \mid N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$f(n \mid N, \nu) = \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1-\frac{\nu}{N}\right)^{N-n}$$

  - in limit of $N \to \infty$:

$$\frac{N!}{(N-n)!} \to N^n \qquad \left(1-\frac{\nu}{N}\right)^N \to e^{-\nu} \qquad \left(1-\frac{\nu}{N}\right)^{-n} \to 1$$

  - Giving:

$$f(n \mid \nu) = \nu^n e^{-\nu} / n!$$

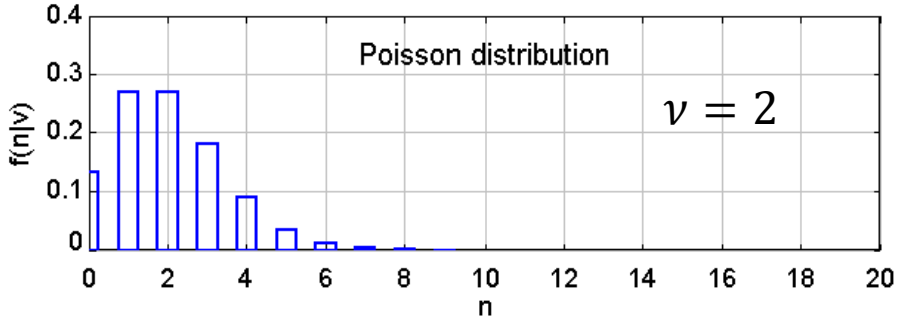
Properties of the Poisson distribution

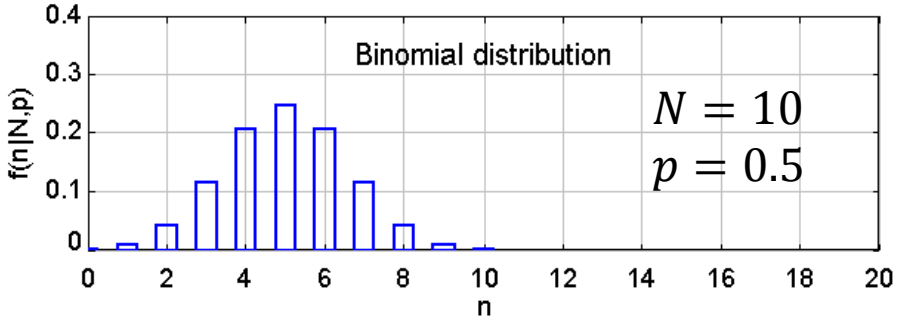
- Expectation value and variance:

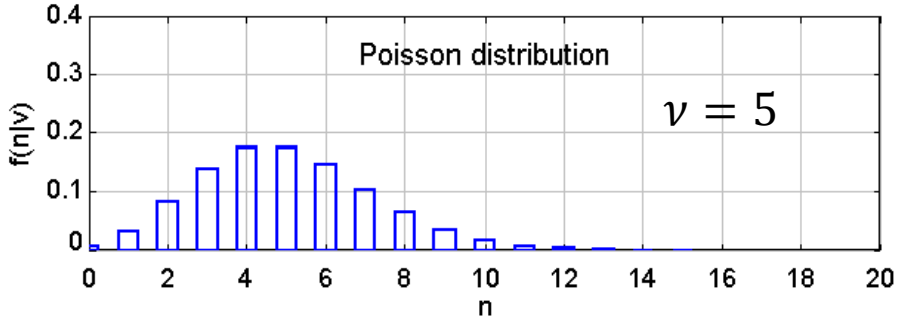
$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n}{n!} e^{-\nu} = \nu$$

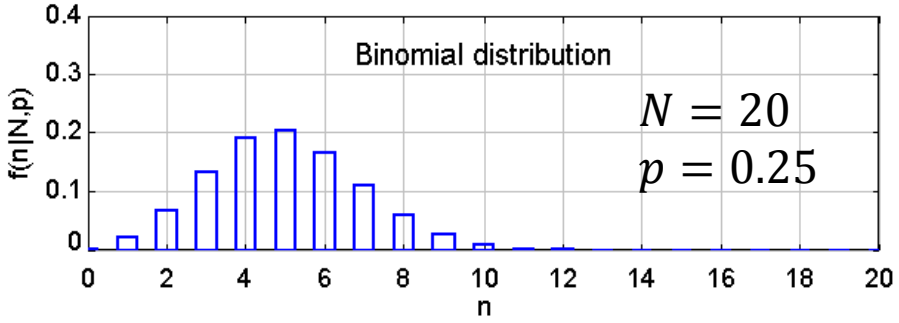
$$V[n] = \sum_{n=0}^{\infty} (n-\nu)^2 \frac{\nu^n}{n!} e^{-\nu} = \nu$$

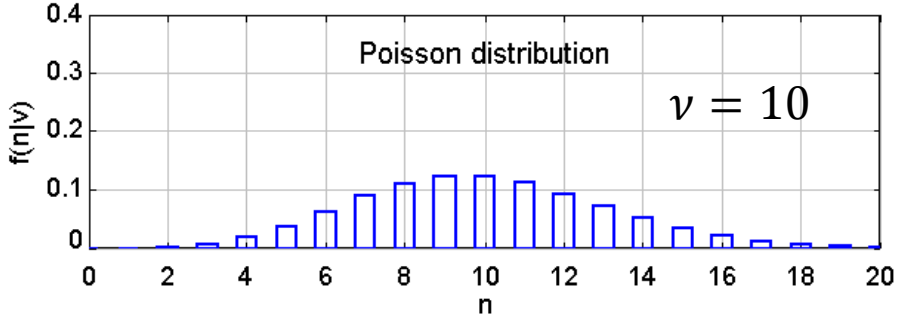
Poisson examples:

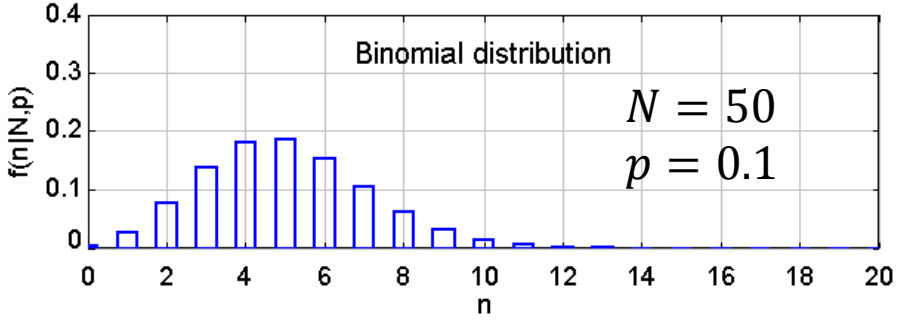






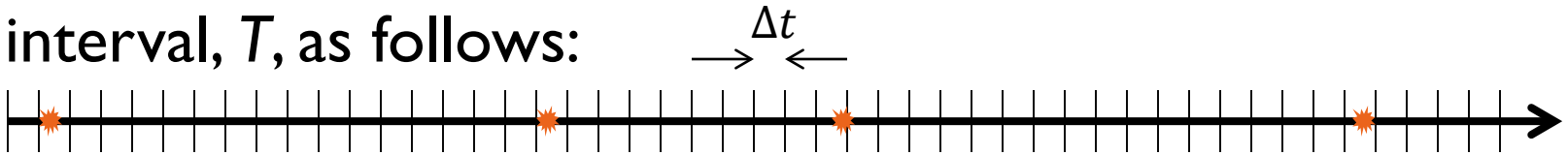






Poisson example: Radioactive decay

- Construct a model of an experiment that counts the number of decays from a radioactive source in a time interval, *T*, as follows:



  - in a short time interval, Δt, assume that the probability for a decay to occur is given by $p = \alpha\, \Delta t$
    - this assumes that the total number of decays is much less than the total number of radioactive nuclei in the source
  - this can be considered to be *N* observations, $T = N\, \Delta t$
    - Let the number of time intervals containing a decay be *n*
    - *n* is a binomial random variable, with pmf: $f(n \mid N, p)$
  - In the limit $N \to \infty$, *n* will be the number of decays, and is a Poisson random variable with pmf: $f(n \mid \alpha T)$

Uses of Poisson distribution

- For any system, where the probability for an occurrence can be assumed to be independent of time, the number of such occurrences observed in a time interval can be modeled by a Poisson random variable
- Possible examples (using simplified models):
  - number of email messages received in a day
  - number of car batteries sold per week in Victoria
  - number of jackpot winners in a casino per week
  - number of homicides in Canada each year

Question: radiobiology

- In a radiobiology research project, a large number of cells are irradiated *in vitro* and the number of surviving cells are counted
  - Suppose the goal of the project is to ascertain whether cell survival depends on the temperature. Measurements would be repeated at different temperatures.
  - The experiment needs to be setup so that under identical conditions, the standard deviations in the surviving cell fractions is about 2%
  - If typically about 40% of cells survive, how many cells need to be irradiated in each experiment?

- Hint: use a model to answer this question!

Question: airport security

- A terrorist tries to bring through airport security a well shielded radioactive source that emits neutrons. The shielding is such that, on average, only 2 neutrons escape per second. The airport security has a neutron detector which detects neutrons, but with an efficiency of only 10% and packages pass through the detector for only 3 seconds.

  - Develop a model for this situation. In this model, what is the probability that the terrorist will be caught?

Question: particle detector

- Consider an experiment that counts radioactive decays from a source, using a new kind of detector
- Suppose you want to understand the behaviour of the detector under different operating voltages. You would like your tests to be sensitive to changes of about 5% in the observed count rate.

  - Develop a model for this experiment. Assume the count rate is about 10 per second. If the standard deviation for the number of counts in the model is to be about 2%, how much time is needed for each voltage setting?

Question: one dice

- Consider a model of an experiment that roles a single dice (a die) 100 times.
  - The die has 6 sides, with numbers from 1-6.
  - What is the expectation value?
  - What is the variance and correlation between the number of 3s and the number of 4s to be observed?
  - Why is this not zero? Why is it negative?
  - Can you change the setup so that the number of 3s and the number of 4s would not be correlated?

Question: three coins

- Suppose you have three coins. They are identical: one side is labelled 0 and the other side is labelled 1. The probability for a coin to show either side after being flipped is equal.
  - What is the probability distribution for the sum of three coins being flipped?
  - Do you use the binomial or multinomial distribution?

Question: two dice

- What is the probability distribution for the sum of two dice being thrown?

Question: Automobile traffic

- The number of cars that cross the Golden Gate Bridge each year is shown here:

Fiscal Year	Avg. Vehicles per day	Total Vehicles	Toll Revenue
2002-2003	106,456	38,856,556	$ 79,427,334
2003-2004	106,234	38,881,684	$ 84,419,500
2004-2005	106,292	38,796,706	$ 84,213,058
2005-2006	106,719	38,952,378	$ 84,746,887
2006-2007	108,263	39,516,006	$ 84,970,839
2007-2008	107,541	39,359,932	$ 85,416,488
2008-2009	104,474	38,132,812	$ 97,121,446
2009-2010	106,784	38,976,078	$100,568,913
2010-2011	110,113	40,191,124	$100,779,715

- Is the variation from one year to the next similar to what would be expected if the crossing probability was constant?

Question: Free throw record

- Ted St. Martin holds the world record for consecutive free throws: an incredible 5,221

  (in 7 hours and 20 minutes)
- Model each throw as being independent with the probability p of making the shot.

  - If the success rate is 99.9% what is the probability to make 5,221 successful free throws in 5,221 attempts?

Question: Baseball

- In 1961 the Philadelphia Phillies lost won only 47 games out of 154. (There are no ties in MLB).
- Furthermore they lost 23 games in a row, a modern day record.

  - How unlikely is it that such a poor team would have such a long losing streak?

Question: Radioactive sources

- Suppose your lab has purchased two types of radioactive sources. Type A has an activity of 5 counts per minute and Type B has an activity of 10 counts per minute. You have 18 of type A and 2 of Type B, but there are no markings to distinguish them and they were put into the same box.
- You select one of the sources and put it inside a perfectly efficient radiation counter for 1 minute. It records 10 decays.
  - What is the probability that the source you selected is Type B?

Question: Bag of coins

- You are given a bag with an unknown number of coins – you only know that there is an odd number of them.
- All the coins in the bag are thrown in the air, and you are told that exactly 3 landed showing heads.

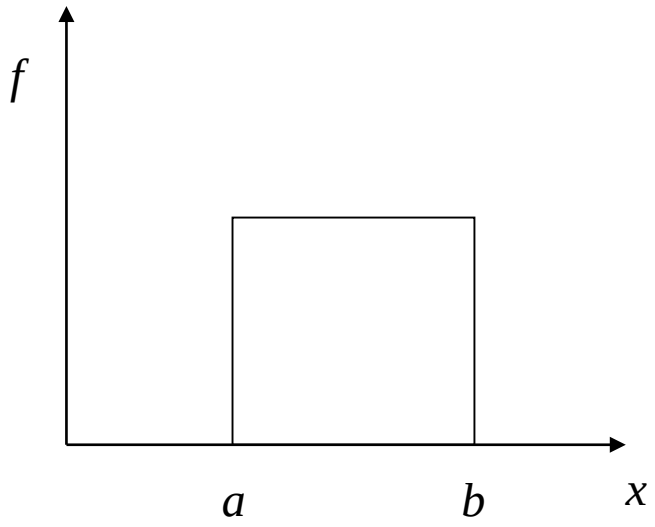
  - What is the most probable number of coins that were in the bag?

  - Repeat the problem, assuming that the prior probability for the number of coins, N, is proportional to $1/N$

Uniform distribution

- The uniform probability density function is constant between two endpoints, and zero elsewhere.

$$f(x \mid a,b) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



$$E[X] = \tfrac{1}{2}(a+b) \qquad V[X] = \tfrac{1}{12}(b-a)^2$$

Exponential distribution

- Used for modelling a system in which the probability for an occurrence is independent of time
  - The length of time that passes while waiting for the occurrence, is a random variable described by the exponential distribution.
- Example: Lifetime of a particle. Divide the time interval t into n equal subintervals. The probability of a decay to occur in any subinterval is $\alpha\, t/n$. The probability for the particle to remain after time t is

$$\left(1-\alpha\frac{t}{n}\right)^{n} \underset{n\to\infty}{\to} e^{-\alpha t}$$

Exponential distribution

- The p.d.f. for the decay time random variable *T* is derived as follows:

$$P(T < t) = 1 - e^{-\alpha t}$$

$$P(T < t + dt) = 1 - e^{-\alpha (t+dt)} = P(T < t) + P(t < T < t + dt)$$

$$P(t < T < t + dt) = P(T < t + dt) - P(T < t)$$

$$= e^{-\alpha t} - e^{-\alpha (t+dt)} = \alpha e^{-\alpha t} dt = f(t \mid \alpha)\, dt$$

- Usually the pdf is expressed as follows:

$$f(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau} \quad \text{for } t \geq 0$$

- Properties:

$$E[T] = \tau \quad V[T] = \tau^2$$

Gaussian (normal) distribution

- This distribution has many applications.
  - Central limit theorem: The sum of n independent random variables X_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean μ and variance σ^2 in the limit that n approaches infinity, where

$$\mu = \sum_{i=1}^{n} \mu_i \qquad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$

    - regardless of the forms of the individual p.d.f.s for X_i (under fairly general conditions)

Gaussian distribution

- The pdf is given by

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

$$E[X] = \mu$$

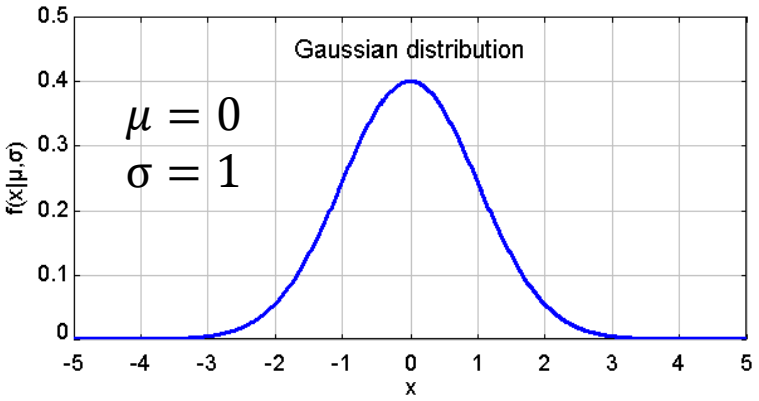
$$V[X] = \sigma^2$$

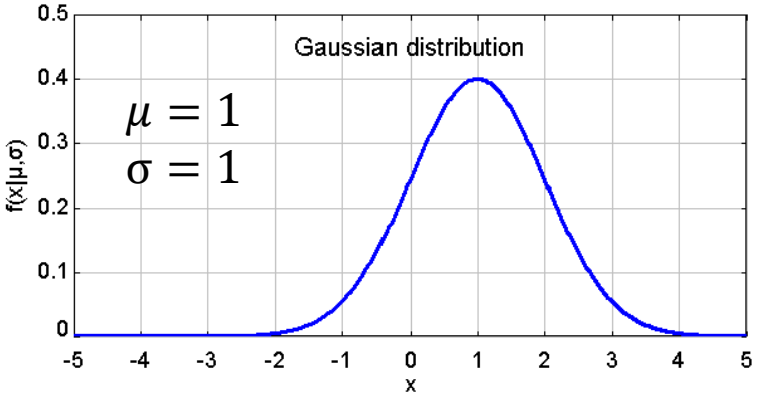
  - standard Gaussian:

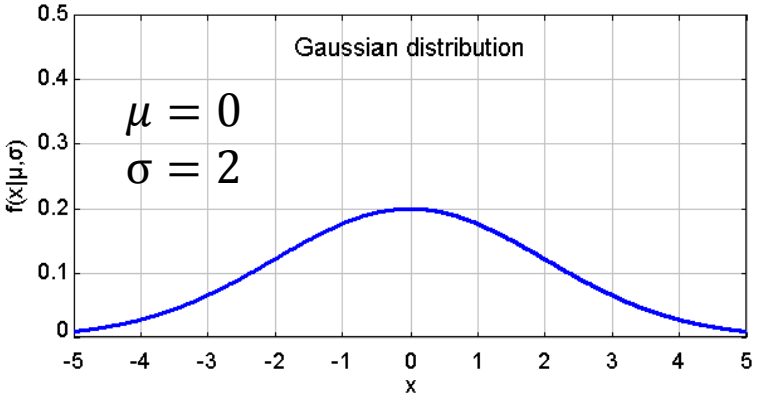
$$\varphi(x) = f(x \mid 0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2} \right)$$

    - cumulative:

$$\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}} \right) \right)$$





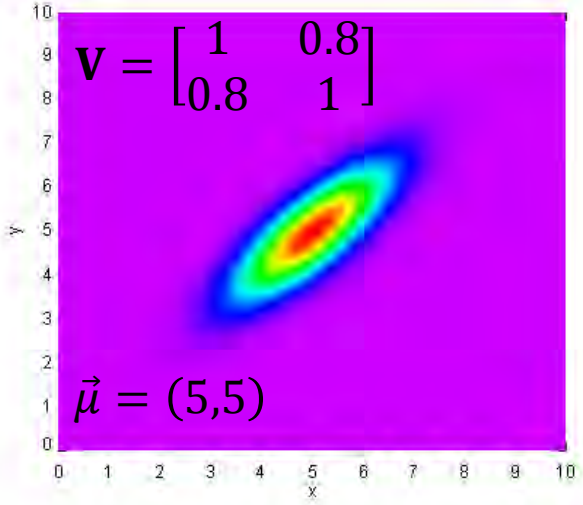


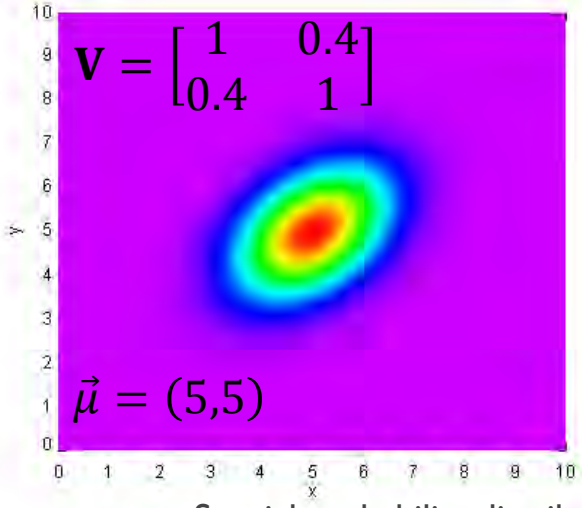
multi-dimensional Gaussian

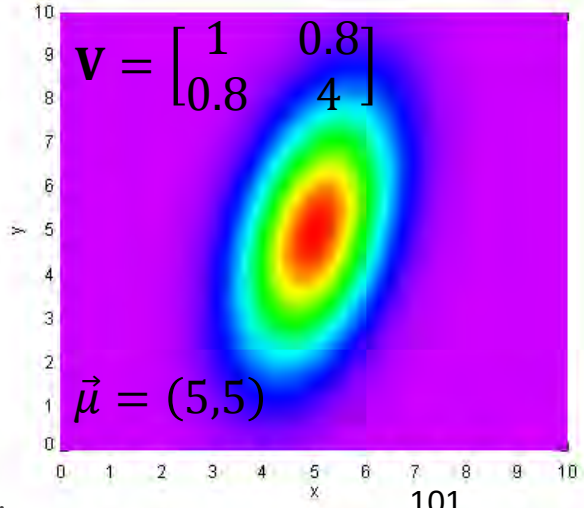
- The joint probability density of N random variables, each distributed according to the Gaussian distribution is given by:

$$f(\vec{x}|\vec{\mu},\mathbf{V}) = (2\pi)^{-N/2}\,|\mathbf{V}|^{-1/2}\,\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\mathbf{V}^{-1}(\vec{x}-\vec{\mu})\right]$$

means — covariance — determinant — inverse







Chi-square distribution (χ^2)

- This distribution describes the pdf of the random variable formed by adding squares of n standard Gaussian random variables:

$$f(z \mid n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

$$\text{for } z \equiv \chi^2 \geq 0$$

$$\Gamma(n) = (n-1)!$$

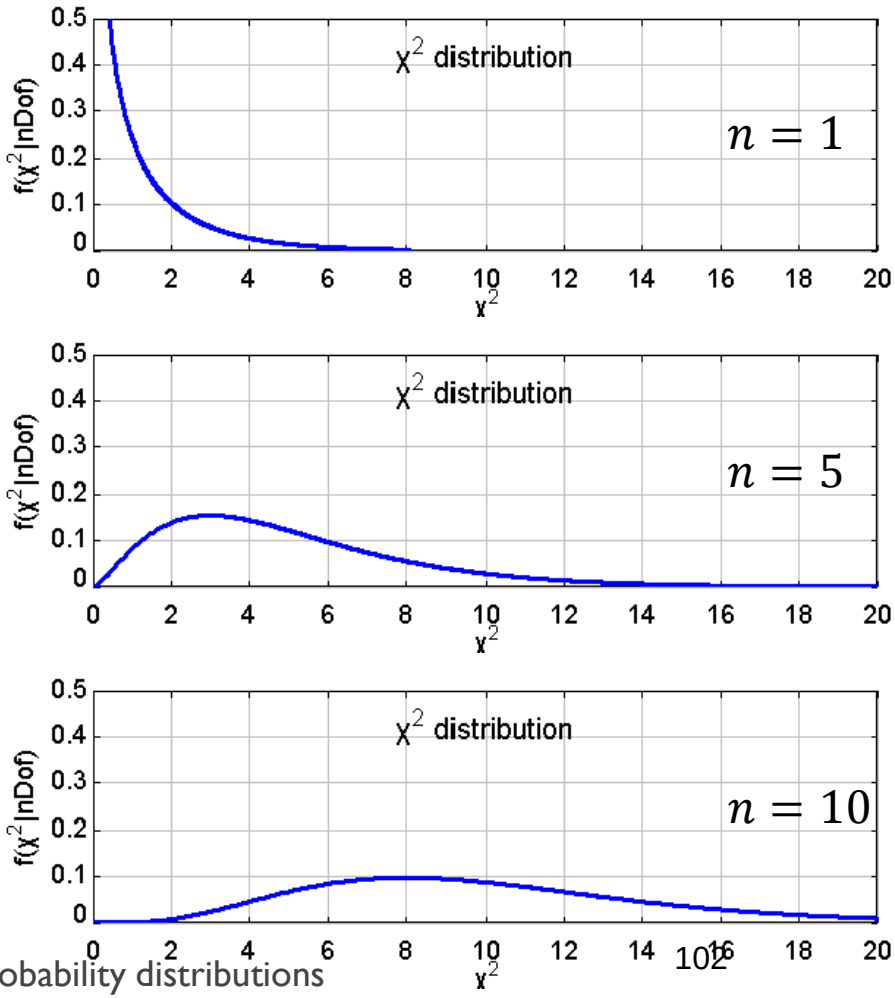
$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1/2) = \sqrt{\pi}$$

$$E[Z] = n$$

$$V[Z] = 2n$$

  - n is called the "number of degrees of freedom"



Question: uniform distribution

- Suppose the random variable, X, is distributed according to the uniform distribution, $f(x|1,2)$
  - What is E[X] ?
  - What is E[1/X] ?

Question: lifetime

- At rest, muons have a lifetime of 2.2 μs
  - What is the meaning of this statement?
  - Is this related to the half-life?
- Suppose we capture a muon in a special trap.
  - According to our model, what is the probability that the muon survives in the trap for 10 μs?
  - Does it matter how much time the muon spent before entering the trap?

Question: weigh scales

- Suppose readings of a scale are well modelled by a random variable distributed according to a Gaussian distribution about the true mass with a standard deviation of 0.02 kg.
- Three calibration masses are put on the scale, with the following results:

Calibrated mass	Measured mass
1 kg	1.02 kg
2 kg	1.94 kg
3 kg	2.96 kg

  - What is the probability that the sum of the squares of the differences would be as far away from the true values (or further), according to this model.

Question: 2 lifetimes

- Suppose the unstable particle A is produced in an accelerator experiment. A has a mean lifetime of τ_A and decays into the unstable particle B. B has a mean lifetime of τ_B and decays into the stable particle C.
  - What is the probability density for the elapsed time between the creation of particle A and the creation of particle C?

Questions: χ^2 distribution

- Derive the form of the χ^2 distribution for $n = 1$
- What is the probability that the sum of the squares of two random variables, each described by the standard Gaussian, will be less than 2?
- What is the probability distribution for the sum of two random variables that each follow a χ^2 distribution for $n = 2$?

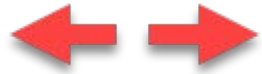
Question: multi-dimensional Gaussian

- Consider the form of the multi-dimensional Gaussian when the covariance is

$$V = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

  - Show that this form can be rewritten as

$$f(x, y) = f(x|\mu, \sigma = 1)\, f(y|\mu, \sigma = 2)$$



Monte Carlo methods

D. Karlen / University of Victoria and TRIUMF

Monte Carlo Methods

- To describe unpredictable behaviour, we develop mathematical models that use the concept of probability and random variables.

- Monte Carlo methods are computer implementations of these models, in which the outcomes of random variables are assigned by numerically generated sequences
  - these sequences are generally called "random numbers"

Random numbers

- What are random numbers?
  - suppose R is a random variable described by the probability density function, $f(r)$
    - a single (unpredictable) outcome of R is a random number
    - a sequence of numbers, $r_1, r_2, r_3, \ldots, r_n$ are said to be random numbers that are generated according to $f(r)$, iff each number is an independent outcome of R
  - properties of a sequence of random numbers
    - in the limit as n goes to infinity, the fraction of numbers in the sequence that are in the range (a,b) equals the integral of $f(r)$ over that range, for all a and b
    - you will not see a "pattern", since each outcome of R is unpredictable – and independent of one another
    - they are NOT random variables (since the values are known)

Pseudo-random numbers

- Monte Carlo methods use numerically generated values to simulate the outcome of a random variable
  - since a numerical algorithm is used to generate the sequence, the numbers are completely predictable – some people call them "pseudo-random"
  - for good pseudo-random number generators, the pattern is not apparent and it turns out that they behave like truly random numbers for many applications
    - likewise, the system that we are modelling with random variables may actually follow a predictable pattern that we are not aware of!
  - in the following "random number" will generally mean a "pseudo-random number"…

Generating uniform random numbers

- Most algorithms that generate random numbers do so according to the uniform distribution $f_u(r|0,1)$.
  - Such a sequence can be used in turn to generate random numbers according to an arbitrary pdf, f(x).

- Simple algorithm: Multiplicative Linear Congruential
  - pick a multiplier *a*, modulus *m*, and starting value n_0 , generate the integer sequence according to the rule:

$$n_{i+1} = (an_i) \bmod m$$

random numbers generated according to $f_u(r|0,1)$ are given by:

$$r_i = n_i / m$$

ejs_mlc.jar

Multiplicative Linear Congruential method

- The choice of multiplier, modulus, and seed are made so that the period is as long as possible
  - once a sequence starts repeating, this cannot be considered a good approximation of a random sequence
  - for 32 bit integers, the maximum period is the number of all integers that can be represented, about 2×10^9
  - this length can be realized, for example
    - a = 40692
    - m = 2147483399
  - other choices can lead to short periods or other behaviours…
    - try a = 65539 and m = 2^{31} with applet…
    - change a by +/- 3

Transformation method

- In this method, a function is found, $x = x(r)$, that transforms uniformly distributed random numbers into random numbers that are distributed according to some other pdf, $f(x)$.
- To find the function, set the cumulative distributions to be equal:

$$P(R < r) = P(X < x)$$

$$\int_0^r dr = r = \int_{-\infty}^{x} f(x)dx = F_X(x)$$

$$\text{solve for } x \quad \rightarrow \quad x = F_X^{-1}(r)$$

  - this equation cannot always be solved… the cumulative of the pdf must be invertible

Example: Transformation method

- Generate a random number according to the exponential pdf

$$f(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau} \quad t > 0$$

$$r = \int_0^t \frac{1}{\tau} e^{-t/\tau} dt = 1 - e^{-t/\tau}$$

$$e^{-t/\tau} = 1 - r$$

$$t = -\tau \log(1 - r)$$

  - since r is a uniform random number $(0,1)$, $t = -\tau \log(r)$ will also work

Question: transformation method

- Describe a method that will generate random numbers according to the pdf,

$$f(x) = \begin{cases} \alpha / \sqrt{x} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where α is a constant. What is α?

Question: transformation method

- Describe a method to generate a sequence of random numbers that follow the pdf:

$$f(x) = \begin{cases} a + bx & 0 < x < c \\ 0 & \text{otherwise} \end{cases}$$

- The constants, a, b, c are all positive numbers.

Question: two lifetimes

- Suppose the unstable particle A is produced in an accelerator experiment. A has a mean lifetime of τ_A and decays into the unstable particle B. B has a mean lifetime of τ_B and decays into the stable particle C.
  - Explain how to generate a sequence of random numbers that would follow the probability density for the elapsed time between the creation of particle A and the creation of particle C?

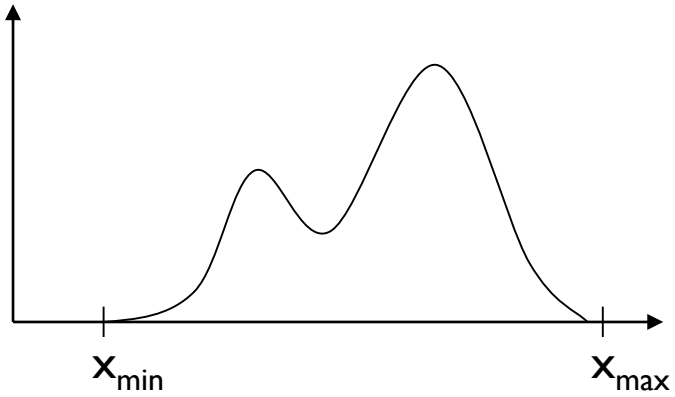
Acceptance-rejection method

- The inversion technique for generating random numbers according to a pdf, works best for relatively simple pdfs, where the cumulative can be inverted analytically
  - In other cases, one could invert the cumulative numerically (root finding) but this is usually inefficient

- A very different approach is to use uniform random numbers to examine the pdf at random locations…

Acceptance-rejection method

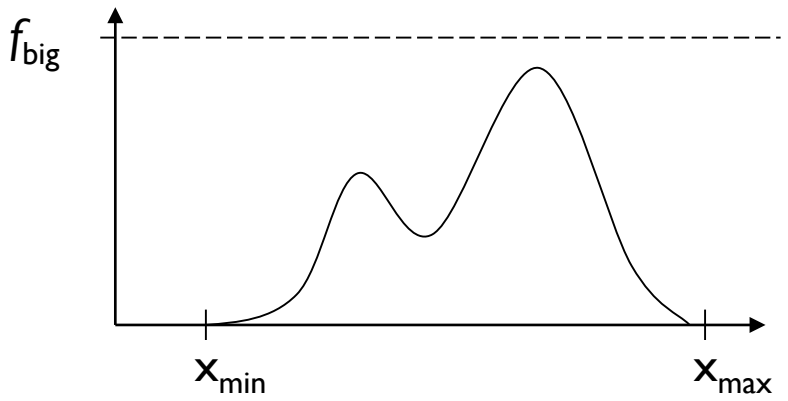
- Example pdf:



  - to span the range of the pdf, spread uniform numbers evenly between x_{min} and x_{max}

$$x_{\text{trial}} = x_{\min} + (x_{\max} - x_{\min})r_1$$

Acceptance-rejection method



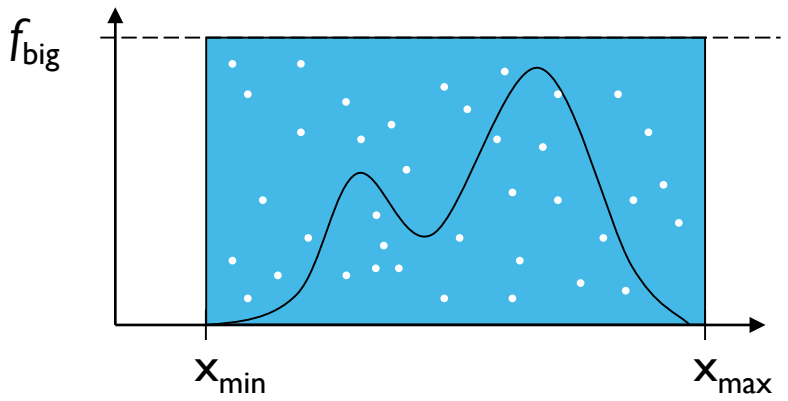
$$x_{trial} = x_{min} + (x_{max} - x_{min})r_1$$

- select a subset of the trial random numbers with a probability proportional to f(x):
  - Accept the trial only if:

$$f(x_{trial}) > f_{big} r_2 \qquad \text{where } f_{big} > f(x) \quad \forall x$$

and r_2 is an independent random number

Acceptance-rejection method

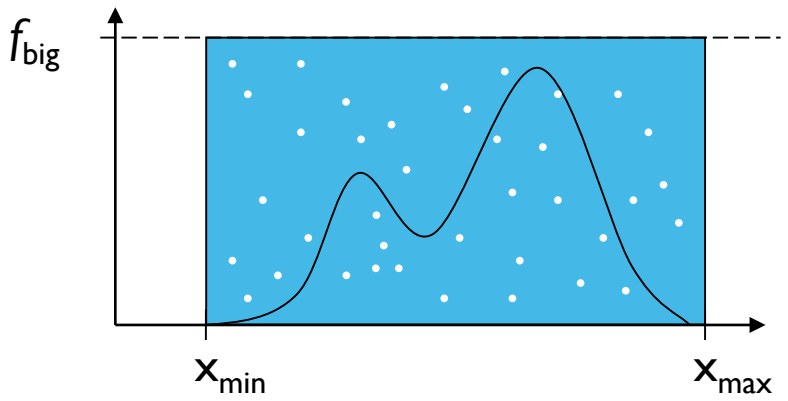


$$x_{trial} = x_{min} + (x_{max} - x_{min})r_1$$

$$f(x_{trial}) > f_{big} r_2$$

- This can be visualized as throwing darts uniformly in the rectangular box above
  - the dots below the curve are accepted

Estimating the integral



$$x_{trial} = x_{min} + (x_{max} - x_{min})r_1$$

$$f(x_{trial}) > f_{big} r_2$$

- The same procedure can be used to estimate the integral of any bounded function
  - the fraction of dots accepted is the fraction of the total area
  - the estimate is an outcome of a random variable K:

$$k = \frac{n_{accept}}{n_{trial}}(x_{max} - x_{min}) f_{big}$$

$$E[K] = \int_{x_{min}}^{x_{max}} f(x)dx = I$$

Estimating the integral

$$k = \frac{n_{\text{accept}}}{n_{\text{trial}}}(x_{\max} - x_{\min}) f_{\text{big}} \qquad E[K] = \int_{x_{\min}}^{x_{\max}} f(x)dx = I$$

- K follows the binomial distribution, where

$$V_{N_{\text{accept}}} = p(1-p)n_{trial} \qquad p = \frac{I}{(x_{\max} - x_{\min}) f_{\text{big}}}$$

  - so, the variance in the estimate for the integral is

$$V[K] = V[N_{\text{accept}}]\left(\frac{(x_{\max} - x_{\min}) f_{\text{big}}}{n_{\text{trial}}}\right)^2 = p(1-p)n_{\text{trial}}\left(\frac{I}{pn_{\text{trial}}}\right)^2 = \frac{1}{n_{\text{trial}}}\frac{1-p}{p}I^2$$

  - the relative uncertainty is

$$\frac{\sigma_I}{I} = \sqrt{\frac{1}{n_{\text{trial}}}\frac{1-p}{p}}$$

Acceptance-rejection method

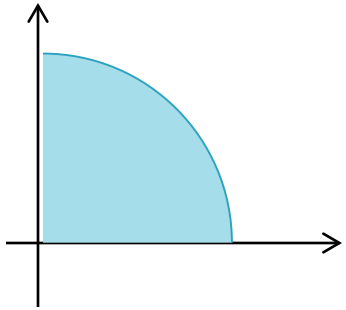
- The acceptance-rejection method is easily generalized for multidimensional pdfs
  - throw darts into a n-volume
- The method is inefficient if pdf has sharp peaks
  - very few trials are accepted
- The method cannot be used if pdf has a pole or defined over an infinite range
  - for example,

$$f(x) = \alpha / \sqrt{x} \quad 0 < x < 1$$

$$f(x) = e^{-x} \quad x > 0$$

Question: Quarter-circle

- Use the general approach of acceptance-rejection to estimate the area of a quarter-circle of radius 1.



- Plot the estimated area as a function of number of trials
- Plot the difference between the estimated area and the true area as a function of the number of trials
  - overlay the expected functional dependence of the standard deviation (see equation for σ_I), earlier.

Importance sampling

- Importance sampling is a hybrid of the two methods (inversion and acceptance-rejection)
  - consider $f(x)$, a complicated pdf for which the acceptance-rejection technique is inefficient or cannot work
  - generate trial random numbers, x_{trial}, but not uniformly over the range. Preferentially sample regions where $f(x)$ is larger
    - To do this, use a simplified approximation of the pdf, $g(x)$ in which the inversion technique applies – the trials are random numbers generated according to $g(x)$
  - select a subset of the trials with probability proportional to the "weight": $w = f(x)/g(x)$

Importance sampling

- Example
  - functions are not normalized!

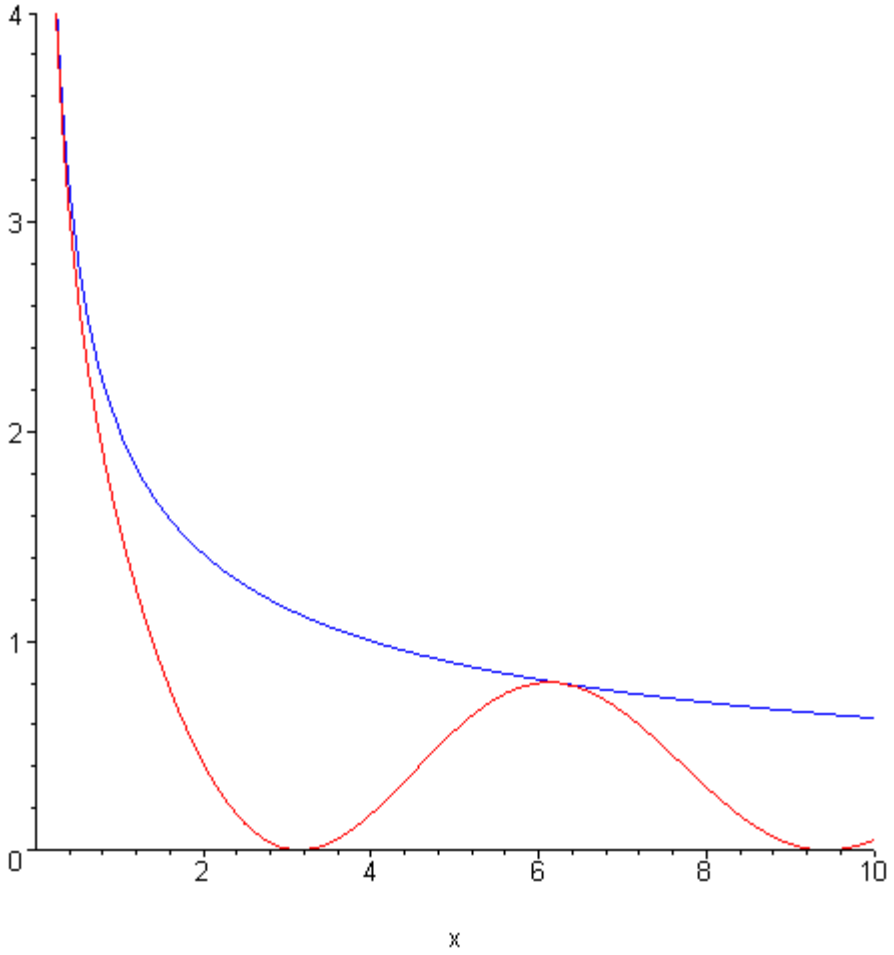
$$f(x) = \frac{1+\cos x}{\sqrt{x}} \qquad 0 < x < 10$$

$$g(x) = \frac{2}{\sqrt{x}} \qquad 0 < x < 10$$

$$\rightarrow \quad x_{\text{trial}} = 10 r_1^2$$

$$w(x) = \frac{f(x)}{g(x)} = \tfrac{1}{2}(1+\cos x)$$

Accept trial if $w(x_{\text{trial}}) > w_{\text{big}} r_2$



Importance sampling

- Notes:
  - in a more complicated example, $f(x)$ might not be analytic – and therefore w_{big} may be unknown
    - choose a value by guessing : if a trial happens to produce a weight larger than this value, then increase w_{big} and start over
  - only the shape of the pdf $f(x)$ is required, not its normalization. The integral can be estimated:

$$k = \frac{n_{\text{accept}}}{n_{\text{trial}}} w_{\text{big}} \int_{x_{\min}}^{x_{\max}} g(x)dx \qquad E[K] = \int_{x_{\min}}^{x_{\max}} f(x)dx$$

$$V[K] = \frac{1}{n_{\text{trial}}} \frac{1-p}{p} E[K]^2 \qquad p = \frac{E[K]}{w_{\text{big}} \int_{x_{\min}}^{x_{\max}} g(x)dx}$$

Importance sampling

- A more efficient estimate of the integral is found by using all trial events (not just those that are accepted):

$$\int_{x_{\min}}^{x_{\max}} f(x)dx = \int_{x_{\min}}^{x_{\max}} \frac{f(x)}{g(x)} g(x)dx = \int_{x_{\min}}^{x_{\max}} w(x)g(x)dx = E[W]\int_{x_{\min}}^{x_{\max}} g(x)dx$$

  - Estimate the expectation value E[W]

$$j = \frac{1}{n_{\text{trial}}} \sum_{i=1}^{n_{\text{trial}}} w_i \qquad V[J] = \frac{V[W]}{n_{\text{trial}}}$$

  - the variance using this approach is always less than the variance from the approach that uses only accepted trials (on the previous page)

Questions: Importance sampling

- Describe methods to generate random numbers that follow the following pdfs:

$$f(x) = e^{-x}\cos^2 x \qquad x > 0$$

$$f(x) = \frac{e^{-x}}{\sqrt{\sin x}} \qquad 0 < x < 1$$

$$f(x) = \frac{e^{-1/x}}{x^2} \qquad x > 0$$

Multidimensional pdfs

- Simple extension for acceptance-rejection method

$$x_{\text{trial}} = x_{\min} + (x_{\max} - x_{\min})r_1$$
$$y_{\text{trial}} = y_{\min} + (y_{\max} - y_{\min})r_2$$
$$f(x_{\text{trial}}, y_{\text{trial}}) > f_{\text{big}} r_3$$

- For the other methods, use:

$$P(X,Y) = P(X \mid Y)P(Y)$$
$$f(x,y) = f(x \mid y)\, f(y)$$

  - generate y according to f(y)
  - use that value of y, and generate x according to f(x|y)

Markov chain Monte Carlo

- The methods described earlier in this section are designed to produce sequences of random numbers, $\vec{x}_i$ which represent independent outcomes of the random variables $\vec{X}$, described by the joint pdf $f(\vec{x})$
  - this is a good model for repeated experimental measurements when the outcome from one trial is independent from other trials

- Markov chain Monte Carlo (MCMC) methods are efficient for producing sequences in large-dimensions, when there is no requirement for them to represent independent outcomes

Markov chain Monte Carlo

- The most common MCMC algorithm is the Metropolis-Hastings method:
  - i is the serial number of the random sequence of points
  - start the sequence ($i = 0$) with an arbitrary point $\vec{x}_0$, such that $f(\vec{x}_0) > 0$
  - consider a proposed next point in the sequence: an outcome of the random variable $\vec{Y}_i$ defined by the pdf $q(\vec{y}|\vec{x}_i)$
    - note that the next point depends on its preceding point
  - accept the proposed next point with probability $\rho(\vec{x}_i, \vec{y}_i)$ and otherwise repeat the current point
    - that is, if accepted: $\vec{x}_{i+1} = \vec{y}_i$ and otherwise $\vec{x}_{i+1} = \vec{x}_i$
    - $\rho(\vec{x}, \vec{y}) = \min\left\{\frac{f(\vec{y})}{f(\vec{x})}\frac{q(\vec{x}|\vec{y})}{q(\vec{y}|\vec{x})}, 1\right\}$
  - repeat the previous two steps, to produce a long sequence of points distributed according to $f(\vec{x})$

Markov chain Monte Carlo

- The initial points in the chain may not be representative of the pdf, and therefore it is common to discard them
  - To decide if you have discarded enough, check if your result changes significantly when you discard fewer or more
  - For high dimension studies, 1000s of points may need to be discarded

- Further simplification: choose a "symmetric" pdf:

$$q(\vec{y}|\vec{x}) = q(\vec{x}|\vec{y})$$

  - then: $\rho(\vec{x}, \vec{y}) = \min\left\{\frac{f(\vec{y})}{f(\vec{x})}, 1\right\}$
    - always jumps to the proposed point if it has higher density

  - Examples: uniform or Gaussian distribution centered on $\vec{x}$

Generating Gaussian random numbers

- Method #1: Use central limit theorem, and add 12 uniform random numbers (0,1) and subtract 6.

- Method #2: Use a variable transformation for the 2 dimensional Gaussian:

$$f(x,y)dxdy = \frac{1}{2\pi}\exp\left(-\tfrac{1}{2}(x^2+y^2)\right)dxdy$$

Generating Gaussian random numbers

- in polar coordinates:

$$f(r,\theta)drd\theta = \frac{1}{2\pi}\exp\left(-\tfrac{1}{2}r^2\right)rdrd\theta$$

$$f(u,\theta)dud\theta = \frac{1}{2\pi}e^{-u}dud\theta$$

- Use the inversion technique to generate u, θ:

$$u = -\log(r_1)$$

$$r = \sqrt{2u}$$

$$\theta = 2\pi\, r_2$$

$$x = r\cos\theta$$

$$y = r\sin\theta$$

A similar method is implemented in java.util.Random:

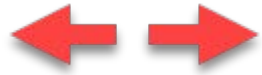
random.nextGaussian()

Question: χ^2 distribution

- Explain methods to generate random numbers that follow the χ^2 distribution
  - Is your method efficient for any number of degrees of freedom?

Question: multi-dimension Gaussian

- Develop a simple method to generate random variables according to a 2-dimensional Gaussian distribution, with standard deviations σ_1, σ_2 and correlation coefficient ρ.
  - Hint: include covariance as a shared variation by combining one-dimensional Gaussian random variables



Testing Hypotheses

D. Karlen / University of Victoria and TRIUMF

Testing Hypotheses

- Measurements can be used to test hypotheses.
  - examples:

Measurement	Hypothesis under test
body weight	diet modification is ineffective
particle ionization rate	particle is a proton
ocean temperatures	climate is unchanging
supernova absorption lines	supernova is type Ia

- Hypothesis under test:
  - sometimes called the "null hypothesis" and labelled H_0
  - simple hypothesis: specific enough to define the probability distribution of the observables (if the hypothesis is true)

Testing Hypotheses

- To test the null hypothesis, properties of alternative hypotheses must also be known:
  - If the probability (density) for the observed value is zero for all alternative hypotheses (while it is non-zero for the null hypothesis) → a definitive statement can be made:

    "The null hypothesis is true."

  - Typically the observed value is possible under null and alternative hypotheses → no definitive statement can be made
    - Instead, a probability statement is made about the null hypothesis
    - The probabilities calculated by Bayesians and Frequentists are very different

Testing Hypotheses: Bayesian

- Calculate the posterior probability that the hypothesis (H_0) is true, given the observed data (x):

$$P(H_0 \mid x) = \frac{P(x \mid H_0)}{P(x)} P(H_0)$$

  - requires a prior probability for the null hypothesis
  - requires all alternative hypotheses and their prior beliefs to be specified:

$$P(x) = P(x \mid H_0)P(H_0) + P(x \mid H_1)P(H_1) + P(x \mid H_2)P(H_2) + \ldots$$

    - can be difficult to compute!

Testing Hypotheses: Bayesian

- If just one alternative hypotheses, consider the "posterior odds":

$$\frac{P(H_0 \mid x)}{P(H_1 \mid x)} = \frac{P(x \mid H_0)}{P(x \mid H_1)} \frac{P(H_0)}{P(H_1)}$$

posterior odds — likelihood ratio — prior odds

- posterior probability:

$$P(H_0 \mid x) = \frac{1}{1 + 1/\text{posterior - odds}}$$

Testing Hypotheses: Frequentist

- Assuming the null hypothesis to be true, calculate the probability of measuring data as "anomalous" (or more anomalous) than observed
  - this quantity is called a "P-value"
    - small values do not lend support to the hypothesis
- Problems with this approach:
  - often misinterpreted as the probability that H_0 is true
  - ad-hoc definition of "more anomalous"
    - normally that which is further from expectation if H_0 was true
    - must specify what is "more anomalous" prior to observing data
  - need to consider the probability of data not observed!
  - more than one way to calculate this probability
    - test statistic and stopping rule must be specified beforehand

Testing Hypotheses: Frequentist

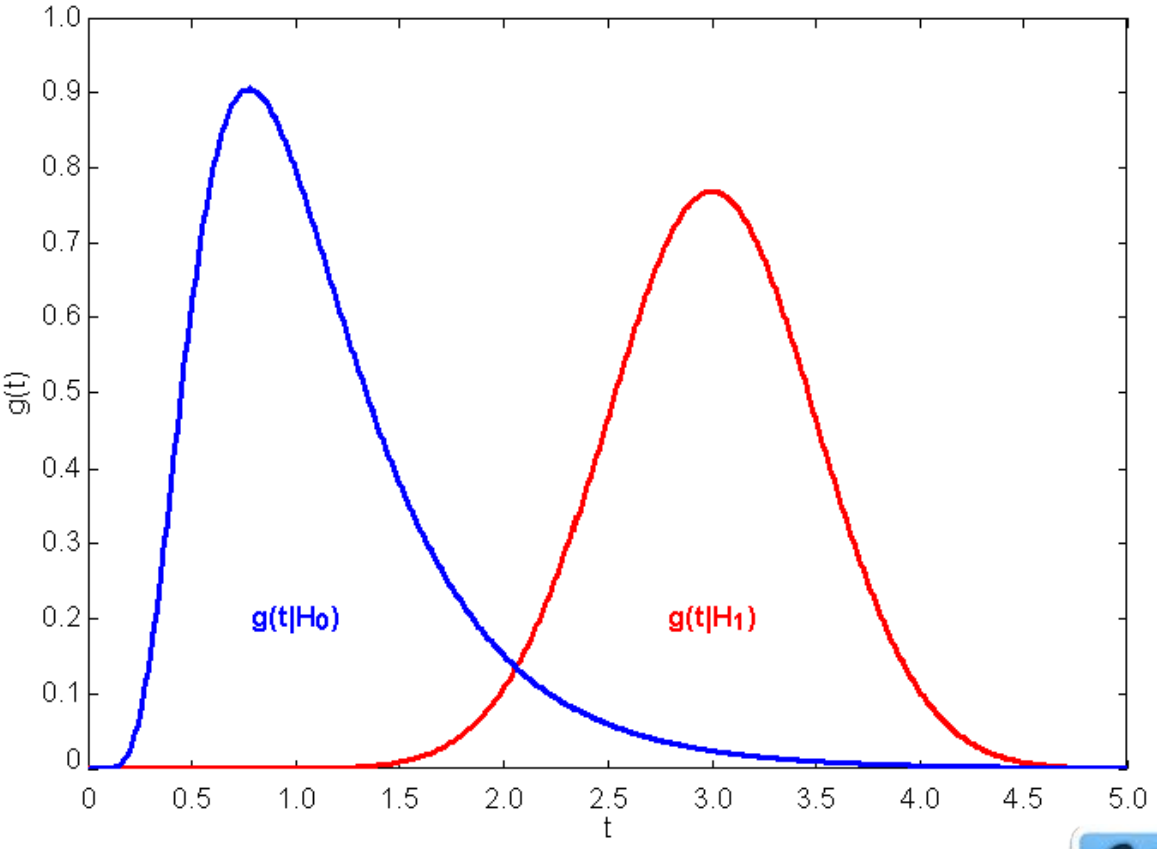
- Test statistic, T
  - a random variable – a function of the random variables that represent the experimental observables
    - designed so that the pdf $g(t|H_0)$ and $g(t|H_i)$ have limited overlap
    - for experiments that measure a single quantity, x, the test statistic can be the observable itself, i.e. T = X. If the measurement is to be repeated it can be the sample mean.
    - for experiments with many observables, sometimes complex functions of the observables are used (defined by artificial neural networks, genetic algorithms, boosted decision trees, etc.)
    - for a single alternative hypothesis, you cannot do better than the likelihood ratio
- Stopping rule
  - how you decide when enough measurements have been made
  - note that g(t|H) depends on the stopping rule

Testing Hypotheses: Frequentist

- Example:
  - only two hypotheses, H_0 and H_1
  - larger values of t are more anomalous
  - report the P-value:

$$p = \int_{t_{obs}}^{\infty} g(t \mid H_0)\, dt$$

ejs_plot5.jar



Testing Hypotheses: Frequentist

- Alternatively, define acceptance and rejection regions (before data is seen)

Significance level:

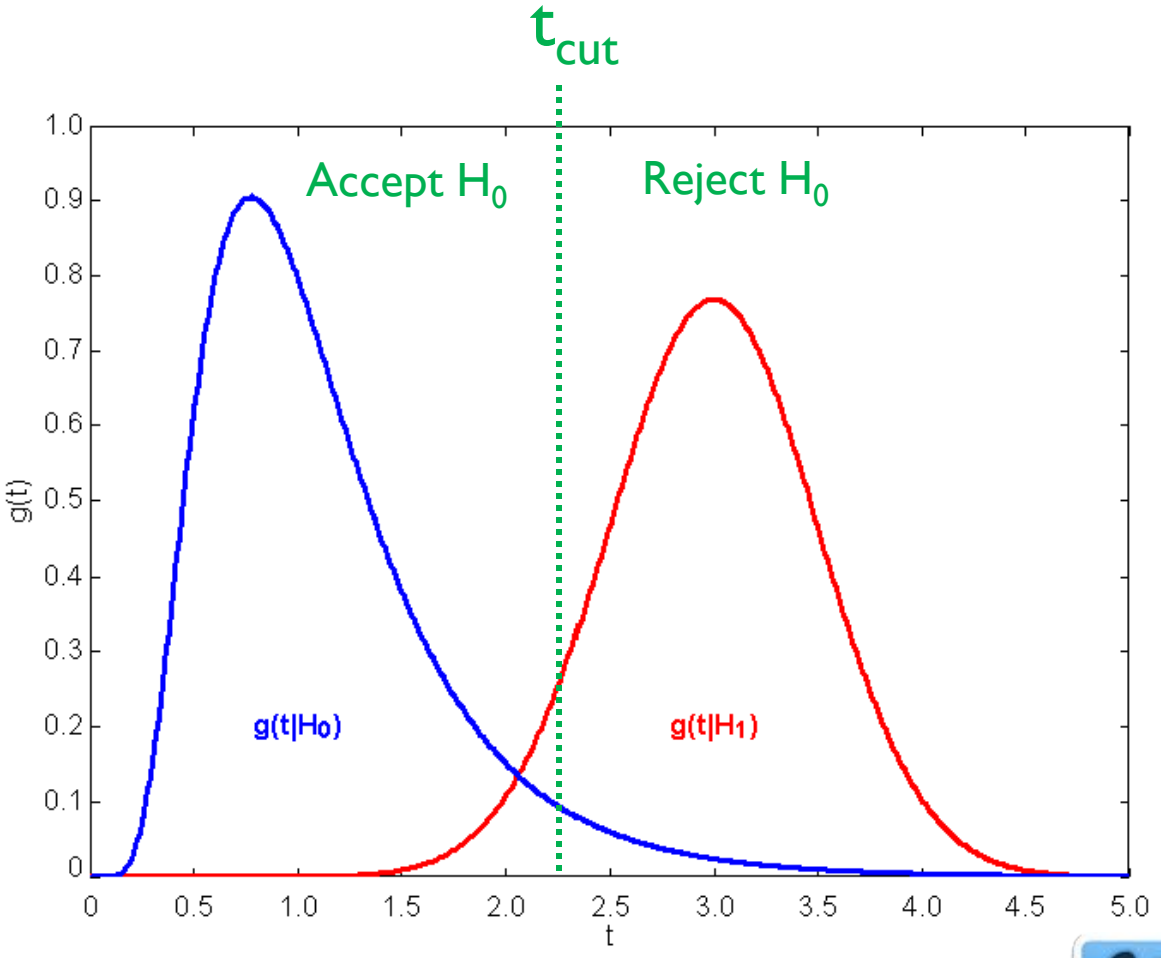
$$\alpha = \int_{t_{cut}}^{\infty} g(t \mid H_0)\, dt$$

probability of Type-I error

Power:

$$1-\beta = \int_{t_{cut}}^{\infty} g(t \mid H_1)\, dt$$

probability of Type-II error



Testing Hypotheses: Frequentist

- Suppose that prior to collecting data, the frequentist decides to set the significance to $\alpha = 0.05$, which corresponds to $t_{cut} = 2.3$
- If the observed result is $t = 2.5$, the frequentist can state: "The experiment rejects the null hypothesis at the 95% confidence level."
  - If the experiment is of public interest, it is not unusual to see newspapers report:

    "Scientists 95% certain that the current theory is wrong!"

    - This is an incorrect statement. Try to come up with a correct statement suitable for the general public!

Questions about P-values

- Is the P-value an outcome of a random variable?
- Describe the probability distribution of P-values if the null hypothesis is correct?
- Describe the probability distribution of P-values if there is only one alternative hypothesis, and it is correct?
- Suppose a large number of experiments perform tests with significance 0.1. If the null hypothesis is true in all cases, what fraction outcomes will reject the null hypothesis at the 90% confidence level?

Question – lifetimes

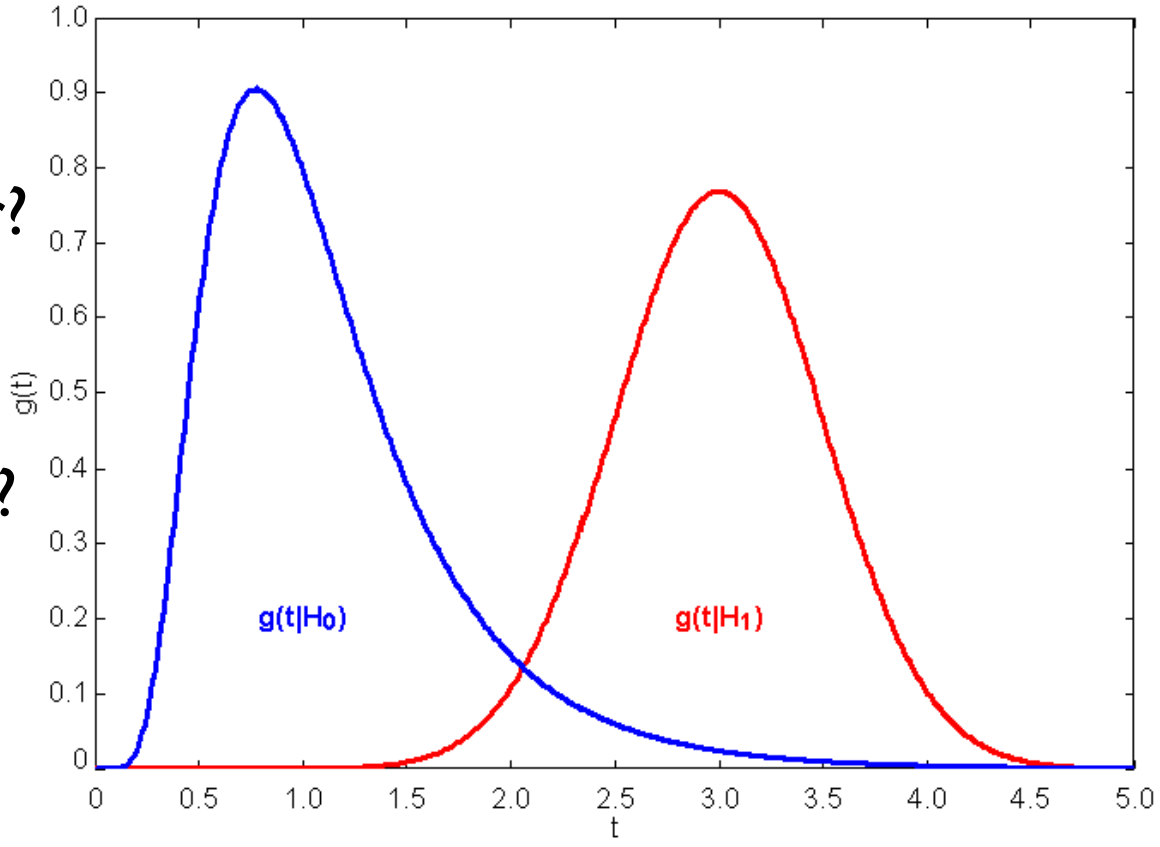
- Suppose an unstable particle was observed to exist for 4 s before it decayed. Hypothesis H_0 states that it was particle χ_0 , with lifetime τ = 1 s. Hypothesis H_1 states that it was particle χ_1, with lifetime τ = 2 s.
- Due to the production processes, the abundance of particle χ_0 is 100 times that of particle χ_1.
  - Describe how a Frequentist and Bayesian would test the null hypothesis (H_0 in this case). What statements would they make?

Questions – Bayesian

- If the prior belief in the two hypotheses was equal, what values of t would give you greater belief in H_0?

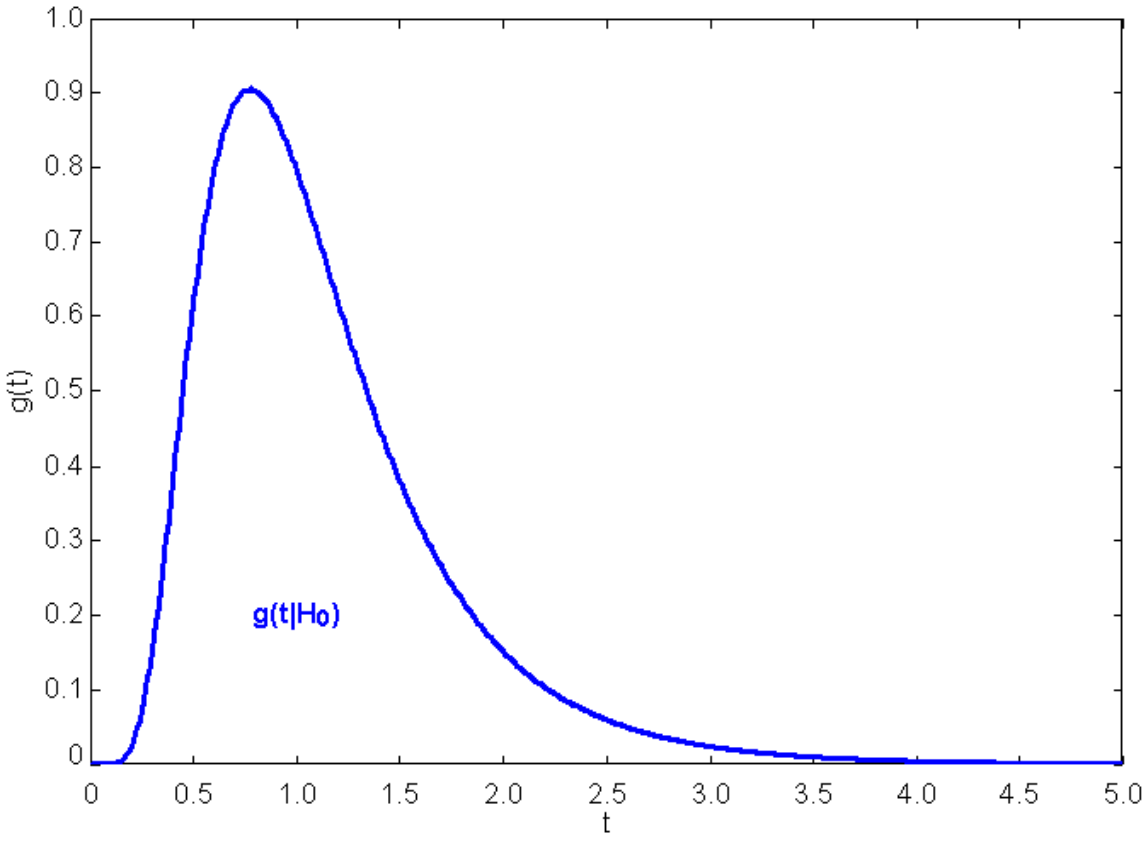
- What about if the prior belief in H_0 was 10 times larger?

- What conclusions are made if t = 0.1 ?



Testing Goodness-of-fit

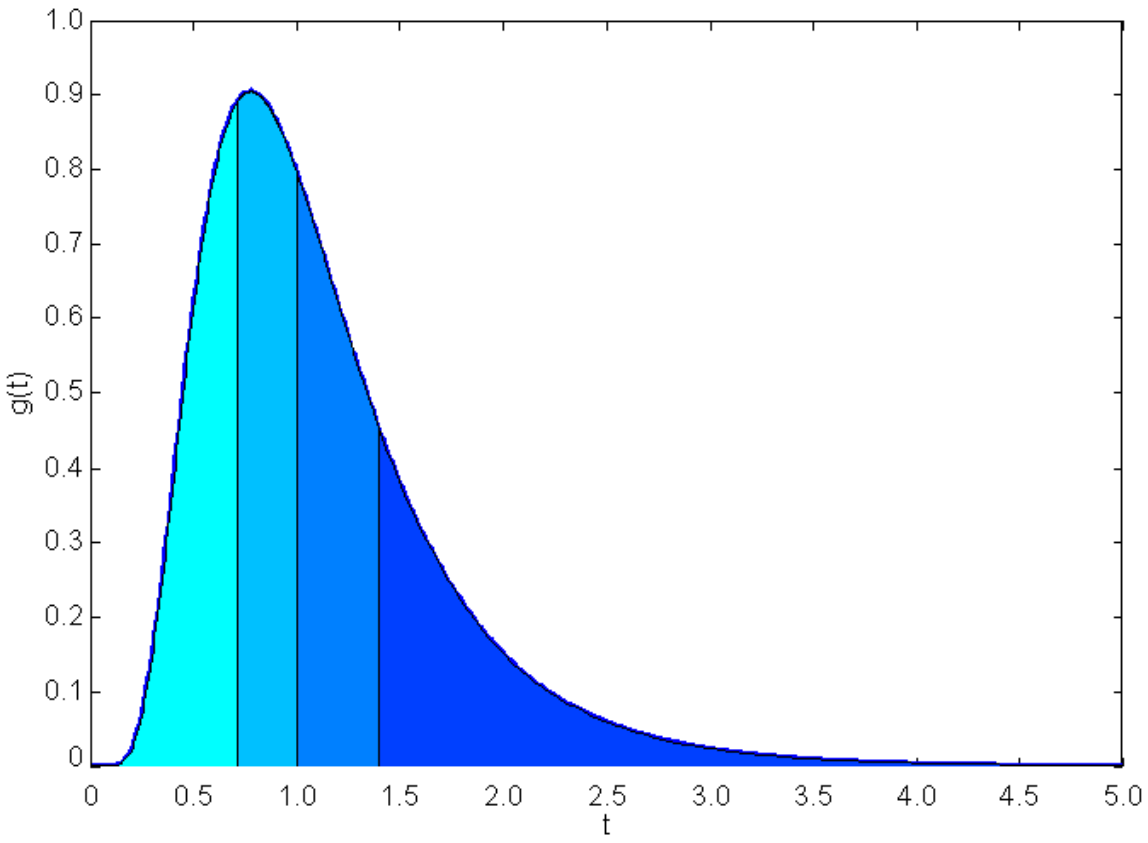
- When alternative hypotheses are ill-defined or not specific enough to define probability distributions, the question still arises whether or not observed data is compatible with the null hypothesis
  - Perform a test of the null hypothesis



Question: How can you define a rejection region?

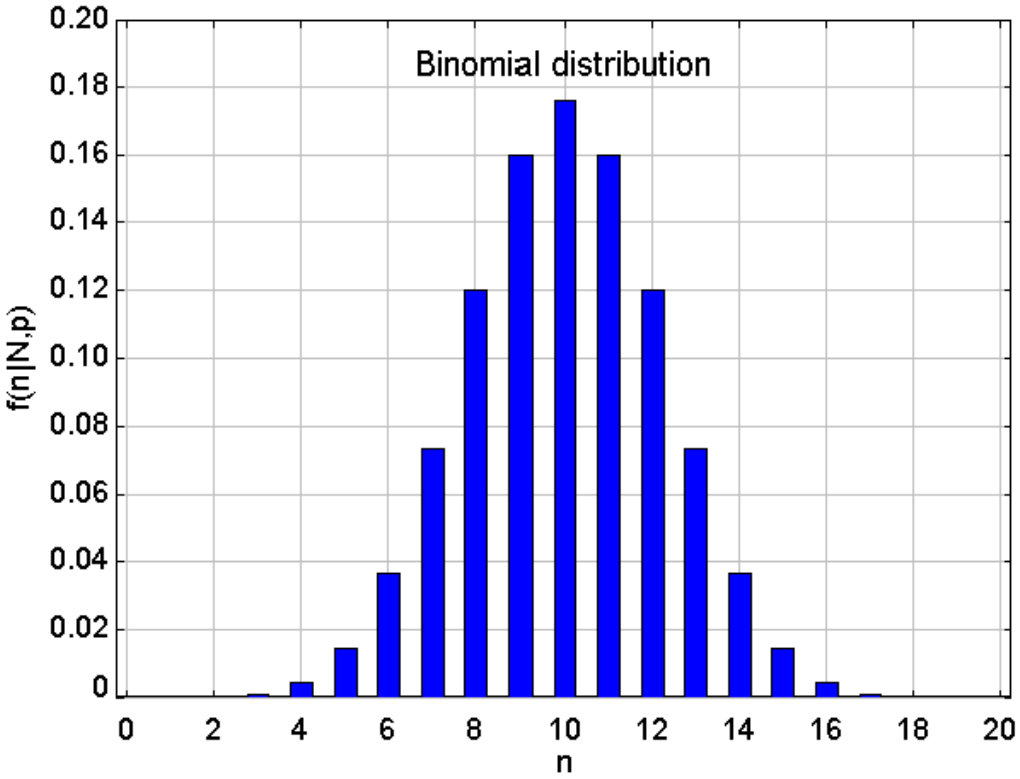
Testing Goodness-of-fit

- It is not reasonable to choose regions of low probability density as rejection regions
  - The 4 regions below have equal probability – why choose one of them to be in the rejection region?
  - You can only choose the rejection region if you have some information about the alternative hypotheses



Testing Goodness-of-fit: Example

- Test the nature of a coin by flipping it 20 times. Null hypothesis: the number of heads should follow the binomial pmf with $p_{head} = 0.5$.
  - Suppose 17 heads are seen. Does this seem anomalous?
  - Calculate the P-value: probability to observe a result as anomalous (or more anomalous)
    - with a small P-value, the data do not support the null hypothesis



Testing Goodness-of-fit: Example (cont.)

- To calculate the P-value, you need to consider the distribution of outcomes for alternative hypotheses:
  - a rigged coin might have $p_{head} > 0.5$
    - larger number of heads than tails would be expected
    - if the rejection region are samples with large number of heads:

$$\text{P}-\text{value} = \sum_{n=17}^{20} f(n|N = 20, p = 0.5) \cong 0.0013$$

  - prior to observing data, alternative hypotheses would also include $p_{head} < 0.5$

$$\text{P}-\text{value} = \sum_{n=0}^{3} f(n|N = 20, p = 0.5) + \sum_{n=17}^{20} f(n|N = 20, p = 0.5) \cong 0.0026$$

  - if you use data to decide what is anomalous, the P-values will not be distributed uniformly – decide this beforehand!

Testing Goodness-of-fit: Example (cont.)

- To calculate the P-value, you must also consider the stopping rule:
  - Suppose all we know is that 20 flips were made and there were 3 tails observed.
  - The previous calculation assumed the flipper decided to stop after making 20 flips. P-value = 0.0026.
  - Suppose the rule was to stop after observing 3 tails. The calculation is now:

$$\text{P}-\text{value} = \sum_{n=0}^{2} f(n|N = 19, p = 0.5) \cong 0.00036$$

Questions – flipping coins

- Suppose you flip a coin from your pocket 20 times and you find the result: (17 heads, 3 tails)
  - Given a P-value of 0.26%, would you be willing to bet 2:1 odds that a second set of 20 flips will yield more heads than tails?
  - Would your answer change if the coin was from the pocket of a magician (or con-artist)?
  - Some might argue that according to the "Law of averages" that with a fair coin, the next set of 20 flips should have less heads than tails. What is wrong with that argument?

Question – drug evaluation

- Suppose a study was performed to determine the effectiveness of a new drug that is designed to improve the chances of a person's body to accept a donated kidney and that the rejection rate for untreated patients is 30%.
  - A total of 10 patients received the drug treatment and only 1 suffered kidney rejection.

- Is there good reason to publish these findings in support of the drug's effectiveness?

Question – Casino criminal

- Suppose a criminal is known to have visited one of the five casinos in town this past weekend
  - The criminal has a tool that allows him to win the special jackpot for any of the dollar slot machines
  - The slot machines are programmed to have a jackpot winner on average every one million games
- During the weekend, the Apple Casino collected $230,000 from the slot machines, and there were two different winners
- Using the Bayesian and Frequentist methods, test the hypothesis that the criminal did not visit the Apple Casino

Calculating the significance of a signal

- Many experiments are designed to search for "signal" events in the presence of known "background" sources.
  - Suppose 2 events were observed with a background expectation of 0.15.
  - Work out the P-value for the hypothesis that there is no source of "signal" events:

$$P(n \geq n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} f(n \mid \nu_b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} f(n \mid \nu_b)$$

$$= 1 - \left(e^{-\nu_b} + \nu_b e^{-\nu_b} \right) = 0.010$$

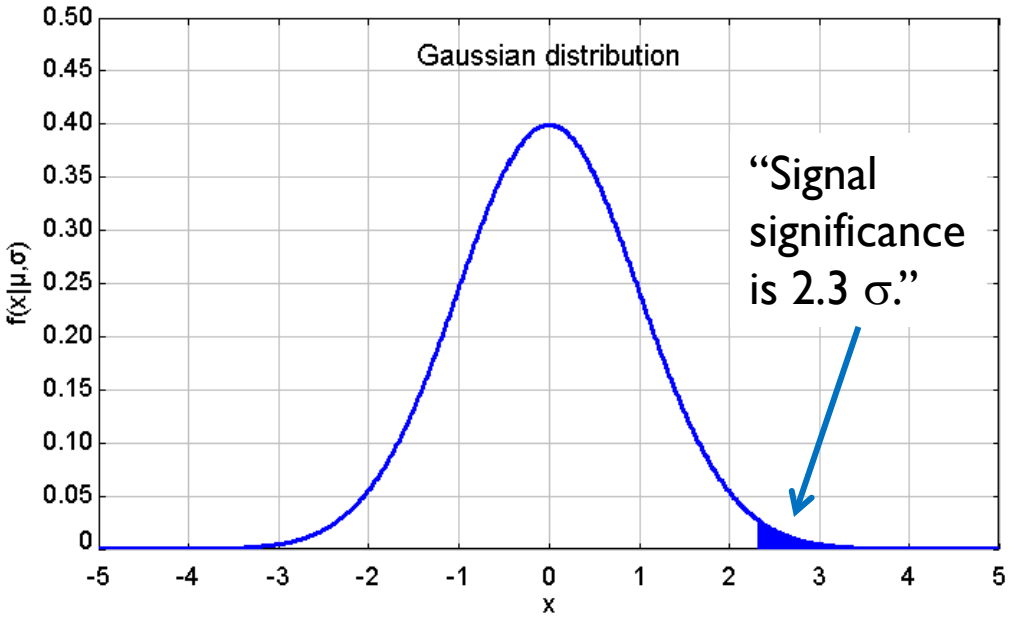
  - Remember: this is not the probability of the no-signal hypothesis!

Calculating the significance of a signal

- It is common to report the significance of a signal with either the P-value or the corresponding number of "sigma" for a Gaussian
  - This corresponds to the value of z, such that:

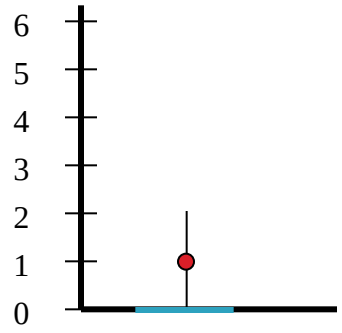
$$\text{P-value} = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

z	P-value
1	0.1587
2	0.0228
2.33	0.010
3	0.0014
4	3.17E-5
5	2.87E-7

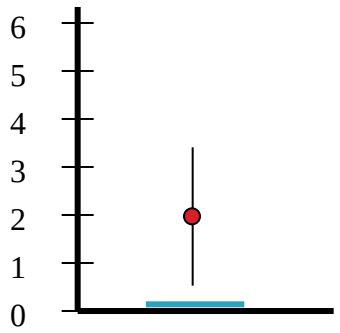


Examples

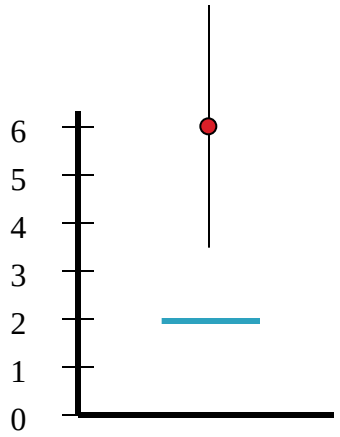
- What are the significances of the signals shown below?
  - "counting data" are presented in a conventional fashion:
    - point indicating the number of observed events (n)
    - an "error bar" size $\pm\sqrt{n}$
    - expected background as a horizontal bar



A: background 0.
observe 1 event



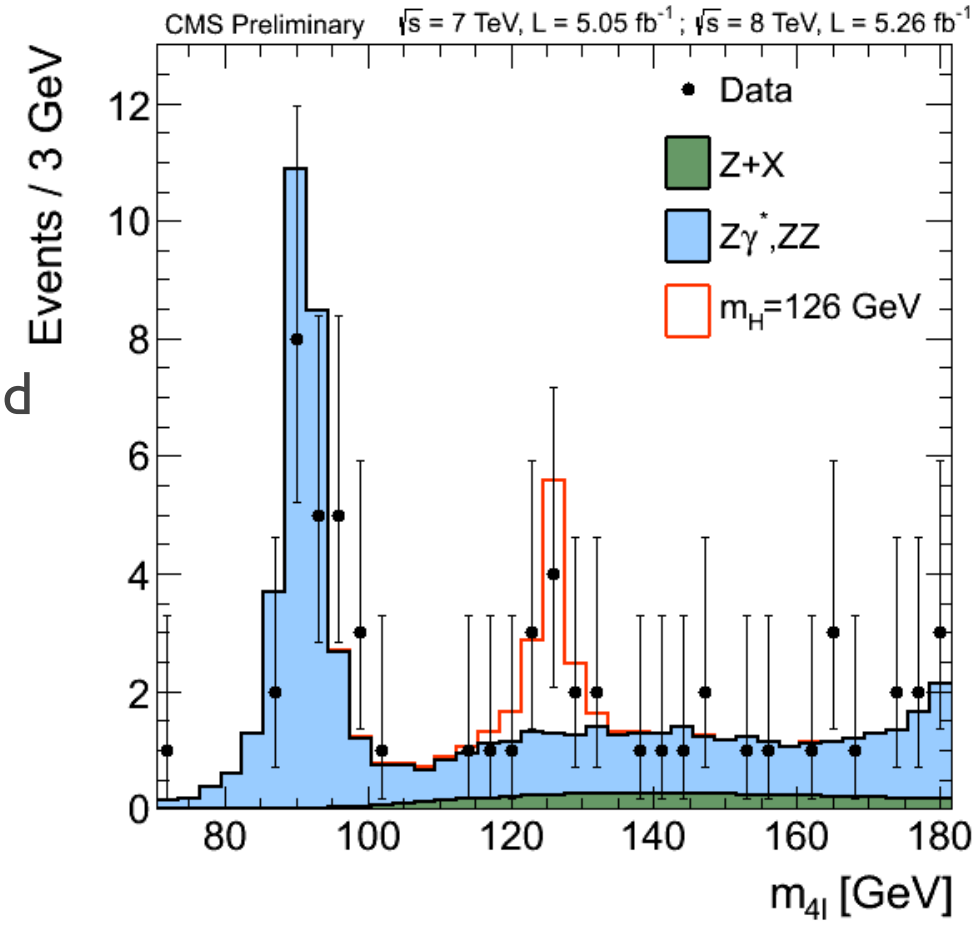
B: background 0.15
observe 2 events



C: background 2.0
observe 6 events

Question – Higgs discovery

- Consider this plot showing an early result of a search for Higgs in the 4 lepton channel by CMS
  - In the range 121-130:
    - 9 observed
    - 4.8 expected background
  - using these numbers, what is the significance?
- Must define method to calculate P-value before observing data.



http://cms.web.cern.ch/news/observation-new-particle-mass-125-gev

Goodness of Fit Methods: Pearson's χ^2 test

- Most common goodness of fit test
- For measurements modelled as outcomes of Gaussian random variables, X_i, with expectation μ_i and standard deviation σ_i, use the test statistic:

$$\chi^2 = \sum_{i=1}^{m} \frac{(X_i - \mu_i)^2}{\sigma_i^2}$$

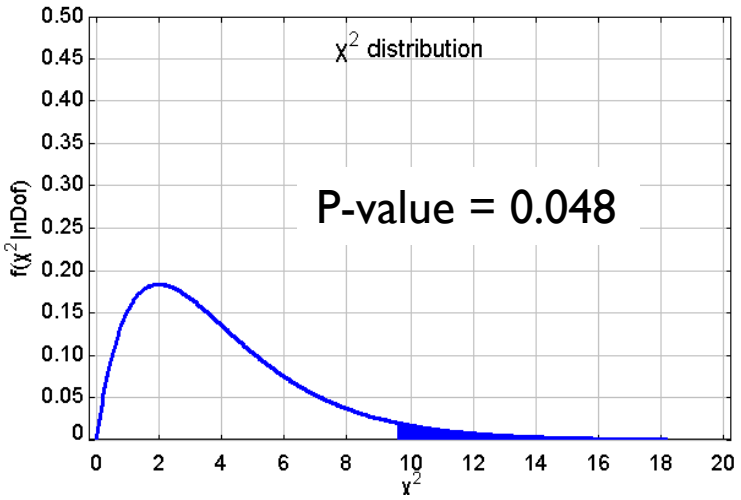
  - under the null hypothesis, χ^2 follows the χ^2 distribution with m degrees of freedom.
  - Alternative hypotheses would yield larger values, so the rejection region (i.e."more anomalous" samples) are those with larger values of χ^2.

Example:

- Measuring the resistance of high resistance resistors is problematic due to the small currents involved. Test the model that the labelled resistances are correct and that measurements have a standard deviation given by,

$$\frac{\sigma(R)}{R} = 0.1\,\frac{R}{10\ \mathrm{G\Omega}}$$

R label (GΩ)	R meas (GΩ)	σ (GΩ)	$(\Delta R/\sigma)^2$
3.00	3.18	0.09	4.0
8.00	7.65	0.64	0.3
10.0	8.5	1.00	2.3
12.0	14.5	1.44	3.0
		Total	**9.6**



Goodness of Fit Methods: Pearson's χ^2 test

- Also applied for histograms (bin contents distributed according to Poisson distributions):
  - provided the expectation for each bin is large enough (roughly $\nu_i > 5$) the test statistic will follow a distribution similar to a χ^2 distribution with m degrees of freedom.

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \nu_i)^2}{\nu_i}$$

- If total number of events is not predicted by model, test the distribution of observations using:
  - in this case the test statistic will roughly follow a χ^2 distribution with $m - 1$ degrees of freedom

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

Goodness of Fit Methods: Student's t-tests

- Testing one Gaussian sample:
  - Pearson's χ^2 test is suitable if the model specifies the expectation and variance.
  - If the variance is unknown, a test of the hypothesis that the sample arises from a Gaussian distribution with known mean μ can be performed by using the test statistic, T

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \qquad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - T follows the Student's t-distribution with $(n-1)$ degrees of freedom (if the hypothesis is true).
  - To calculate the P-value, one must consider if alternative hypotheses would yield outcomes of T which are a greater than zero, less than zero, or either.

Goodness of Fit Methods: Student's t-tests

- Testing two Gaussian samples:
  - A test of the hypothesis that two samples arise from identical Gaussian distributions can be performed by using another test statistic, t

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \text{where} \quad S^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m + n - 2}$$

  - T follows the Student's t-distribution with $(m+n-2)$ degrees of freedom

Goodness of Fit Methods: Student's t-tests

- Testing two Gaussian samples with shared variance:
  - If the two samples are correlated (i.e. paired data have some variance in common) a more powerful test that they arise from identical Gaussian distributions can be performed by using another test statistic, T

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \quad \text{where} \quad {S_d}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \quad D_i = X_i - Y_i \quad \bar{D} = \bar{X} - \bar{Y}$$

  - T follows the Student's t-distribution with $(n-1)$ degrees of freedom

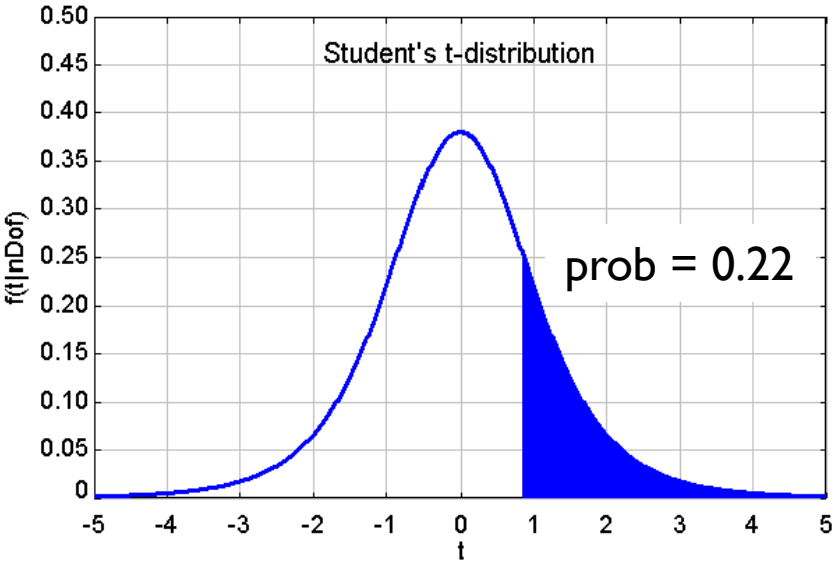
Example

- Suppose that, untreated, the average growth rate of a cancer tumour is 1mm per month. Two treatment regimens, A and B, are performed on a set of patients and the results are tabulated below. Summarize the findings by performing t-tests.
  - Null hypothesis: treatments are ineffective
  - Alternative hypotheses: treatments reduce or increase growth rate

Patient	A rate (mm/mon)	B rate (mm/mon)	A – B (mm/mon)
Alison	0.9	0.8	0.1
Bill	1.3	1.1	0.2
Charlie	0.7	0.6	0.1
David	0.6	0.5	0.1
Elmo	1.2	1.1	0.1
Frank	0.7	0.7	0.0
average	**0.9**	**0.8**	**0.1**

Example: solution

- Test sample A to null hypothesis (μ = 1 mm/mon):
  - $\bar{x} = 0.90 \quad s = 0.29 \quad t = -0.85 \quad \mathrm{nDof} = 5$
  - P-value is probability of more anomalous outcomes:
    $P(T < -0.85 \text{ or } T > 0.85) = 2\,P(T > 0.85)$



P-value = 0.44

- Repeat test for sample B: $t = -1.94 \quad \mathrm{P{-}value} = 0.11$

Example: solution cont.

- Test that sample A and B arise from the same Gaussian distributions:
  - $t = 0.64 \quad \mathrm{nDof} = 10 \quad \mathrm{P{-}value} = 0.54$

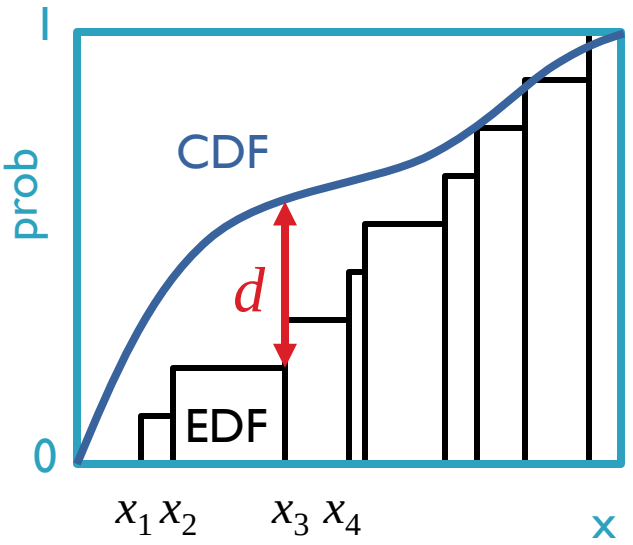
- Tumor growth varies from one person to another. Perform the paired data t-test to compare the two treatments against each other:
  - $t = 3.9 \quad \mathrm{nDof} = 5 \quad \mathrm{P{-}value} = 0.012$

- If hypothesis test significance was set to be 0.05, the null hypothesis is accepted for the first three tests, but rejected for the fourth test (95% CL).

Goodness of Fit Methods: Kolmogorov-Smirnov

- Popular goodness of fit test for the distribution of a single continuous observable
  - Build an "Empirical Distribution Function" (EDF) from the data, and compare it to the Cumulative Distribution Function (CDF) of the null hypothesis
    - Test statistic: the largest difference between the two functions: D
  - The rejection region for the test: larger outcomes of D

- Empirical Distribution Function
  - order the sample, increasing in x_i
  - increment function by $1/n$ after crossing each data value, x_i

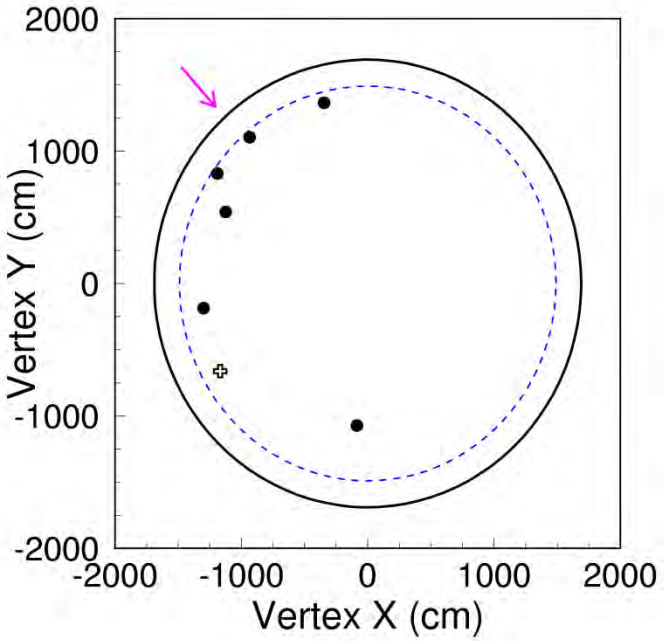


Goodness of Fit Methods: Kolmogorov-Smirnov

- Beneficial features:
  - no binning required and can be used with small sample sizes
  - the test statistic D, under the null hypothesis, only depends on the number of data points
    - $D = \max|EDF(X) - CDF(X)| = \max|EDF(Y) - CDF(Y)|$ where $Y = Y(X)$
- Problematic features:
  - The PDF for D_n cannot be written in a simple analytic form
    - The cumulative distribution for D_n can be computed (all that is needed for P-value)
  - Does not produce P-value=0 when $P(x|H_0) = 0$
  - Forced agreement in tails of observable, so the test is most sensitive in central region

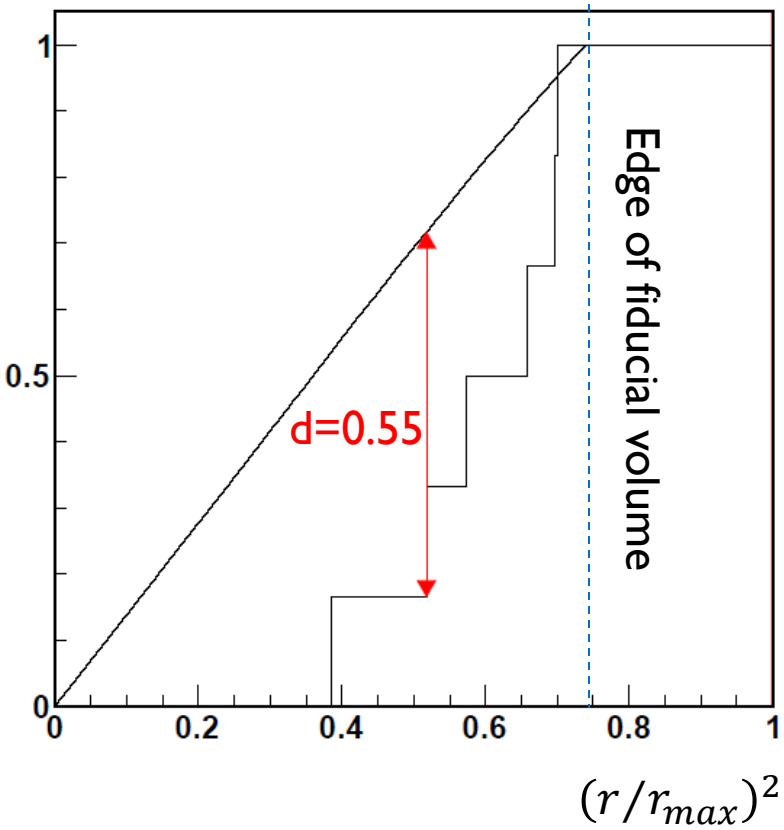
Example

- In 2011, the T2K experiment published the first evidence for the appearance of ν_e interactions in a ν_μ beam.
  - There were only 6 candidate events (with an expected background of 1.5 events) in the cylindrical volume of water that makes up the SuperKamiokande (SK) detector
  - The null hypothesis states that ν_e interactions would have equal probability (and detection efficiency) throughout the so-called fiducial volume of the SK detector. The observed distribution of the locations of the interactions caused some concern.



Example cont.

- Prior to analyzing the data, consistency tests checking for clustering of points along the edge of the fiducial volume were not considered
  - Only after examining the data, was this feature observed and considered of interest
  - A natural check that the distribution of points is consistent with null hypothesis is to use the observable r^2, where r is the cylindrical coordinate, since that observable is uniform in the null hypothesis.



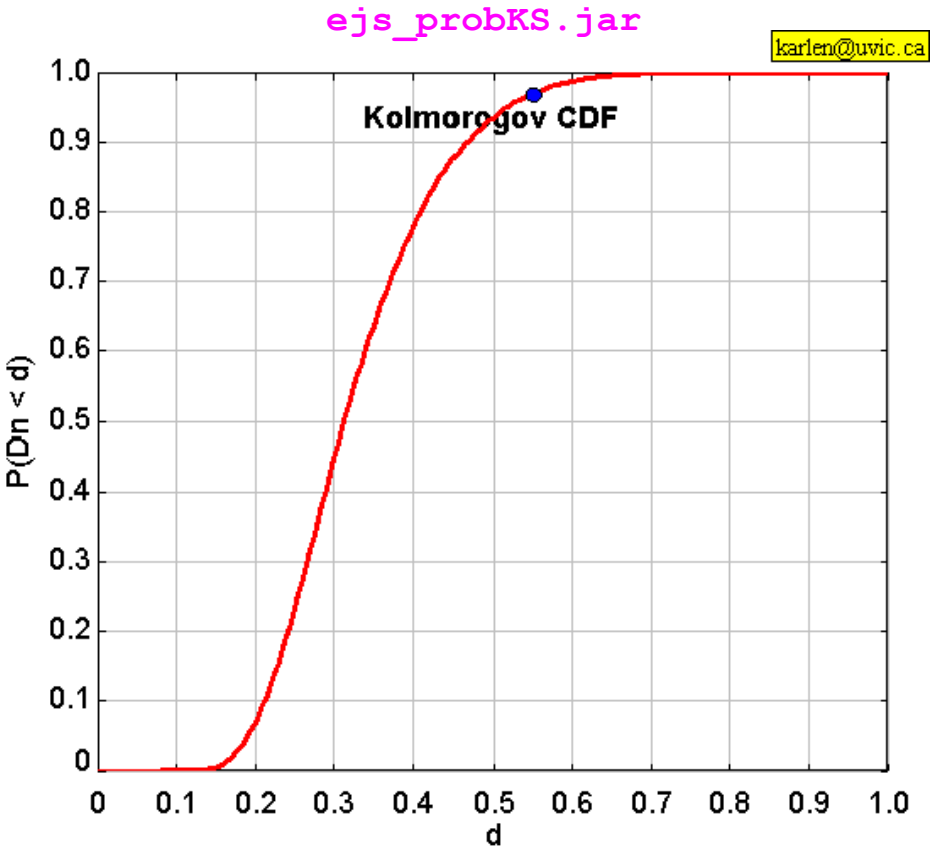
Example cont.

- To calculate the P-value, consider the Kolmogorov cumulative distribution function for n=6:

$$P(D_6 < 0.55) = 0.969$$

P-value = 0.031

- since a large number of unusual patterns could have been observed, not unusual that one has a P-value as small as this
- subsequent data did not show this effect



Question – Test masses

- As part of its quality assurance program, a company that produces calibration blocks for mass scales, weighs each block after it is produced.
  - The company advertises that its 10g blocks are produced with a standard deviation of 10 mg and an expectation of 10.000 g
  - Here are the measurements for the last 5 masses in grams:
    - 10.01 , 10.02 , 10.01 , 10.01, 10.02

- Assuming the scale that is used is properly calibrated, test the hypothesis that the company advertisement is correct

Question – Test masses part II

- Follow up to previous question…
- Suppose the scale used introduces an unknown variation
- Test the hypothesis that the block production has an expectation of 10.000 g.

Question - Lifetimes

- The lifetime of an isotope produced in an experiment is expected to be 1.0 s.
- The first 5 isotopes produced had the following decay times (in seconds)
  - 0.3 , 1.4, 0.9, 0.2 , 2.3

- Perform a goodness of fit test to the null hypothesis

Decision theory

- This section describes techniques used to test a hypothesis with data:
  - Bayesian: calculate the probability that a hypothesis is true
  - Frequentist: calculate the probability of observing data as anomalous (or more) than what was observed, if the hypothesis is true
- These tests summarize the outcomes of experiments, but do not provide enough information to make decisions.

- To make a decision to take action, on the basis of these outcomes, you must also account for the consequences of those actions, for any hypothesis that nature might actually follow.

Decision theory: Two hypotheses

- Simplest case, H_0 and H_1:
  - The outcome of an observable X is x
  - A decision, $d = d(x) = 0 \text{ or } 1,$ is made to act in accordance with H_d
  - A negative consequence would arise if the incorrect hypothesis is chosen. Quantify this by the "Loss function" = $L\,(H_{true}, H_d)$
    - could be a financial cost, loss of health, etc.

- Bayesian: choose the decision rule that minimizes the expected loss, known as the "Risk" $= R = E[L]$

Decision theory: Example

- Consider a Loss Function, $L\ (H_{true}, H_d)$, given by:

$$L(H_0, H_0) = 0 \quad L(H_0, H_1) = \lambda_0$$
$$L(H_1, H_0) = \lambda_1 \quad L(H_1, H_1) = 0$$

- If the decision selects H_0, the expected loss is:
  - $R_0 = \lambda_1 P(H_1|x)$
- Likewise if the decision selects H_1, the expected loss is:
  - $R_1 = \lambda_0 P(H_0|x)$

- Select H_0 if the risk is less: $\lambda_1 P(H_1|x) < \lambda_0 P(H_0|x)$
  - i.e. posterior odds:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)P(H_0)}{P(x|H_1)P(H_1)} > \frac{\lambda_1}{\lambda_0}$$

Question

- You have the job of testing a farm's water supply for an organism that is known to have a long term toxic effect on dairy cattle:
  - The incidence of contamination by the organism is relatively common. About 40% of farms suffer the problem.
  - The test is not definitive. Of all tests that come out positive, 20% are false positives.
  - The test is reasonably efficient, in that 75% of farms with the contamination will have a positive test result.
- There is a treatment that can be added to water supply to kill the organism.
  - If a contaminated water supply is not treated, the milk production will drop by 50%
  - When the treatment is applied, milk production drops by 5% (no matter if the water was contaminated or not.
- What action should be taken if the test comes out positive and if the test comes out negative?

Estimating Parameters & Maximum Likelihood

D. Karlen / University of Victoria and TRIUMF

Parameter Estimation

- Often, theories describing a physical system have one or more undetermined parameters (some refer to such a theory as a composite hypothesis)
- Naturally, the experimental scientist wants to estimate the value of any unknown parameter(s)
  - Suppose an experiment makes n measurements of a single quantity, x. The model of the experiment describes each x as an outcome of a random variable, X, whose pdf is $f(x|\theta)$, where θ is an unknown parameter.
  - The set of n outcomes, called a sample of size n,
    $$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$
    can be used to estimate the value of the parameter θ. A different sample will produce a different estimate: randomness.

Statistics, estimators and estimates

- As before, a statistic is a function of random variables that represent experimental observables
  - When applied for a particular sample, the function returns a single number. That number is also called a statistic.
    - "statistic" refers to both the random variable and the outcome
- An estimator is a statistic used to estimate some property of the pdf, $f(x|\theta)$
  - examples: an estimator of $E[X]$ or of the true value of θ
  - notation: an estimator for θ is written with a hat: $\widehat{\Theta}$
- An estimate is the outcome of $\widehat{\Theta}$ when evaluated with a particular sample
  - an estimate is also written with a hat: $\hat{\theta}$

Properties of estimators

- The estimator is consistent iff it converges to the true value:

$$\lim_{n\to\infty} P(|\hat{\Theta} - \theta| > \varepsilon) = 0 \qquad \forall \varepsilon > 0$$

- The bias of an estimator is given by:

$$b = E[\hat{\Theta}] - \theta$$

  - an unbiased estimator has zero bias for all n

- There is no unique rule for forming an estimator. Usually, estimators are selected which are unbiased and have the least variance

Estimator for expectation value

- Given a sample of size n, the sample mean is an unbiased estimator

$$\hat{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

symbol for average

$$E\left[\hat{\mathrm{M}}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Estimator for variance

- If the expectation value, μ, is unknown, the unbiased estimator for the variance is

$$\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\mathrm{M}}\right)^2 \Rightarrow \hat{\sigma}_X^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)$$

- If μ is known, the unbiased estimator is

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

- Similarly for covariance:

$$\hat{V}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\mathrm{M}}_X\right)\left(Y_i - \hat{\mathrm{M}}_Y\right) = \frac{1}{n-1}\sum_{i=1}^{n} X_i Y_i - \frac{n}{n-1}\hat{\mathrm{M}}_X\hat{\mathrm{M}}_Y$$

$$\Rightarrow \quad \hat{\sigma}_{XY}^2 = \frac{n}{n-1}\left(\overline{xy} - \bar{x}\,\bar{y}\right)$$

Variance of estimators

- The variance can be calculated as usual:

$$\hat{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$V\left[\hat{\mathrm{M}}\right] = E\left[\hat{\mathrm{M}}^2\right] - E\left[\hat{\mathrm{M}}\right]^2 = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{j=1}^{n} X_j\right] - \mu^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} E\left[X_i X_j\right] - \mu^2 = \frac{1}{n^2}\left(n\left(\mu^2 + \sigma^2\right) + \left(n^2 - n\right)\mu^2\right) - \mu^2$$

$$= \frac{\sigma^2}{n}$$

- or more simply by…

$$V\left[X_1 + X_2\right] = V\left[X_1\right] + V\left[X_2\right]$$
$$V\left[aX\right] = a^2 V\left[X\right]$$
$$\Rightarrow \quad V\left[\hat{M}\right] = \frac{\sigma^2}{n}$$

Questions: Estimators

- Is the following estimator for expectation value:

$$\hat{Q} = \frac{1}{n+10}\sum_{i=1}^{n} X_i$$

  - Consistent?
  - Unbiased?

- Show that the estimator for variance, when μ is unknown, is unbiased.

Question:

- The Monte Carlo transformation method is applied below with a set of input numbers that are decidedly not random…

r	uniform	expon $\tau = 1$	Gaussian $\mu = 0, \sigma = 1$
0.1	0.1	0.11	-1.28
0.3	0.3	0.36	-0.52
0.5	0.5	0.69	0.00
0.7	0.7	1.20	0.52
0.9	0.9	2.30	1.28

  - Compare the estimates for the expectation and variance with the true values
  - To estimate the properties of pdfs, is it better to use a sample generated by random numbers or a "grid" of numbers?

Question: uniform distribution

- Produce 100 uniform random numbers $(0 - 1)$
- Calculate the sample mean
- Repeat this 50 times and calculate the variance of the sample means

- Compare this with the expected variance of the estimator

Programming hints

- When using a computer program to calculate the estimates for expectation value and variance, there is no need to
  - store all values in an array
  - loop over the data twice

- instead, keep running sums of
  - n, x, y, x^2, y^2, xy

Maximum likelihood

- A general and powerful method for parameter estimation

  - Again, consider an experiment that measures a single quantity, x, modelled by a random variable, X, that follows the pdf, $f(x|\theta)$. Repeated measurements give the n outcomes, $x_i \ldots x_n$.

  - According to the model, the probability for the outcomes of the random variables to be in the ranges $x_i < X_i < x_i + dx_i$ is

$$\prod_{i=1}^{n} f(x_i \mid \theta)\, dx_i$$

    - expect this to be larger for the true value of θ as compared to a parameter far from the true value

Likelihood

- The likelihood is defined to be

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

  - in frequentist statistics, the likelihood is considered to be a function of the parameter alone. NOTE: It is not a probability density in θ

- Numerically it is usually more convenient to calculate the logarithm of the likelihood function (the so called "log-likelihood"):

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

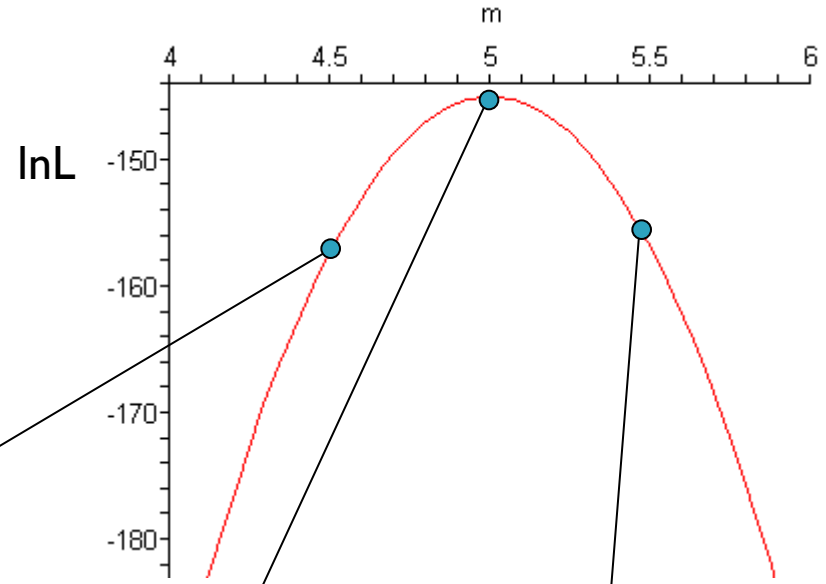
Maximum likelihood estimator

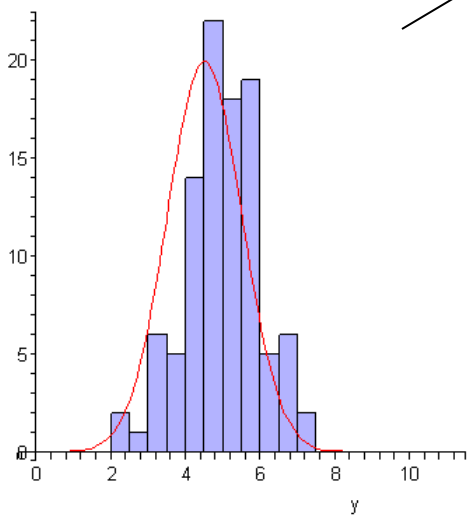
- The maximum likelihood estimate for the parameter is that value of the parameter that maximizes the likelihood (or log-likelihood)
  - the estimate $\hat{\theta}$ can be considered as an outcome of the random variable $\hat{\Theta}$ (the estimator)

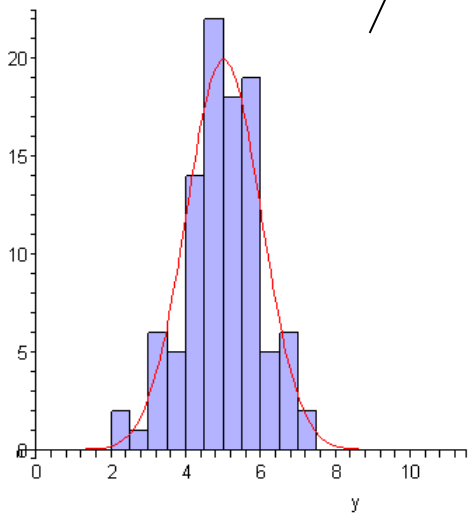
- It turns out that the maximum likelihood estimators usually have good properties: unbiased, minimum variance

Illustration of maximum likelihood

- Consider a Gaussian(5,1) sample of size 100
  - adjust the mean…







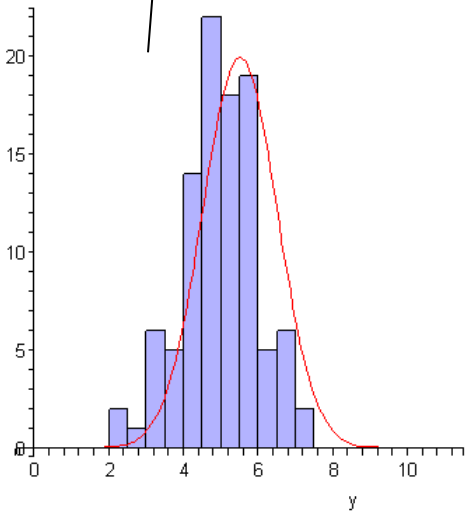


Illustration of maximum likelihood

- Adjust the standard deviation…

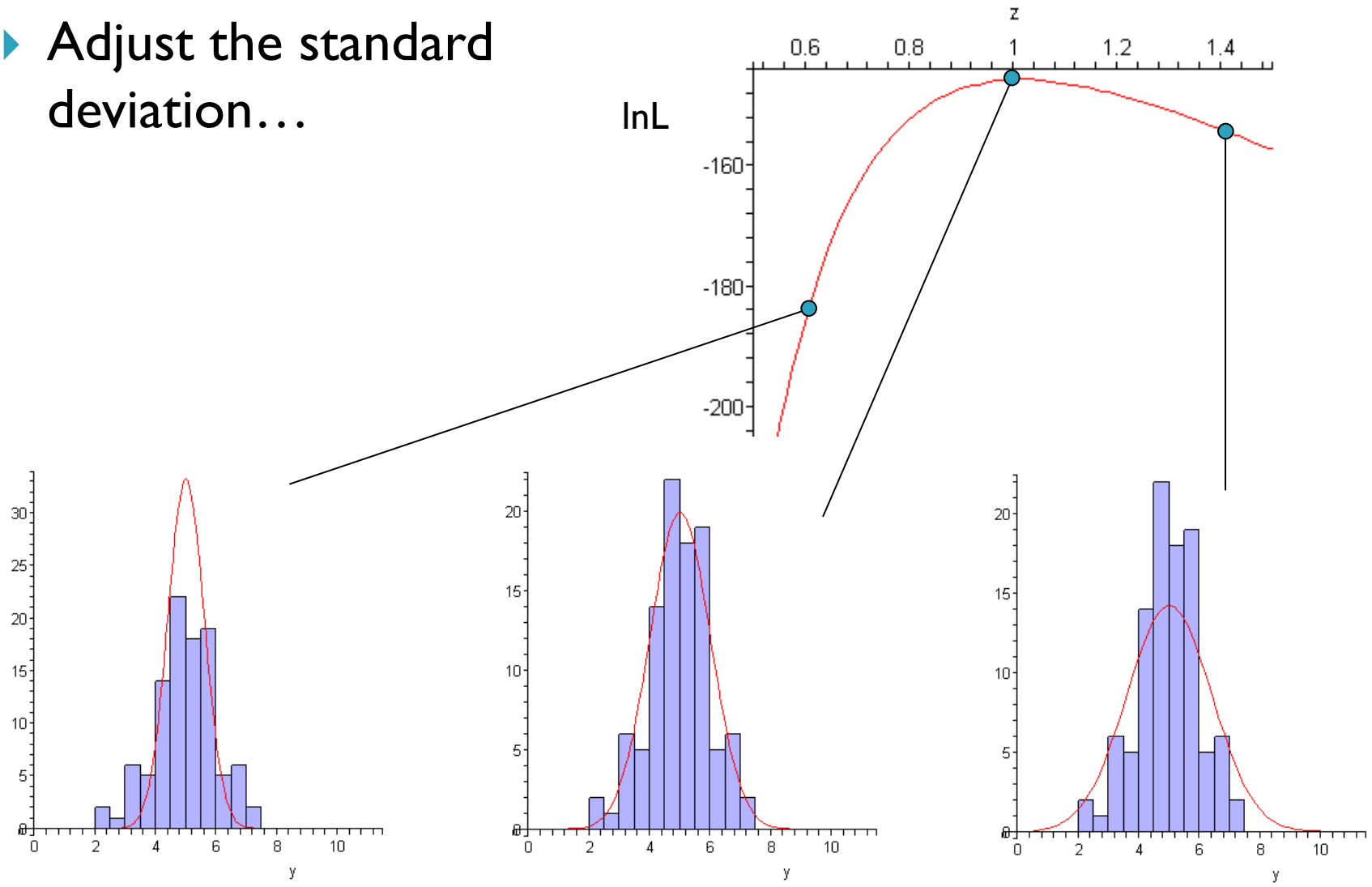
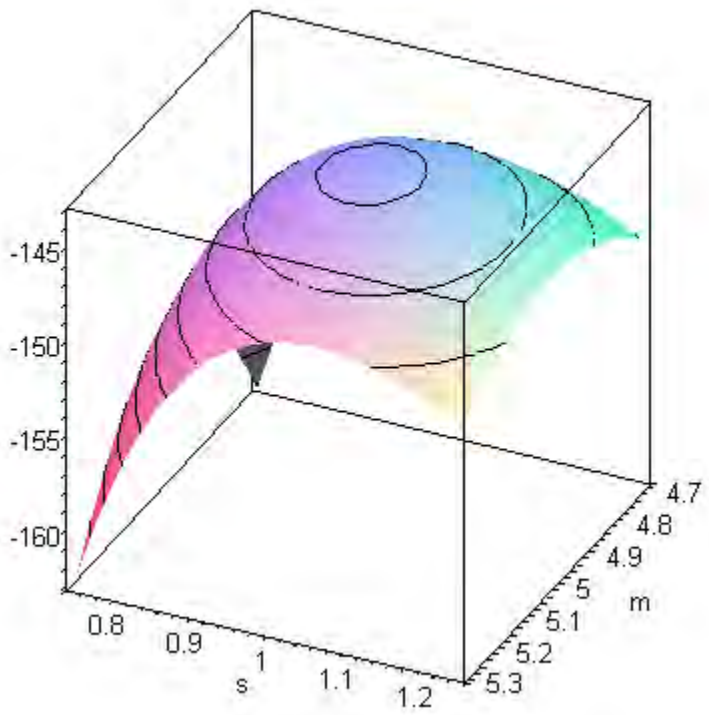


Illustration of maximum likelihood

- The likelihood as a function of mean and standard deviation…

$\ln L(\mu, \sigma)$



Example of an analytic ML estimator:

- Consider an experiment measuring decay times of unstable particles at rest. This can be modelled by a random variable T that follows the exponential pdf:

$$f(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

  - The log-likelihood function is therefore:

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i \mid \tau) = \sum_{i=1}^{n} \left(-\ln \tau - t_i / \tau \right)$$

$$\frac{\partial \ln L(\tau)}{\partial \tau} = \sum_{i=1}^{n} \left(-1/\tau + t_i / \tau^2 \right) = \frac{1}{\tau^2} \left(-n\tau + \sum_{i=1}^{n} t_i \right)$$

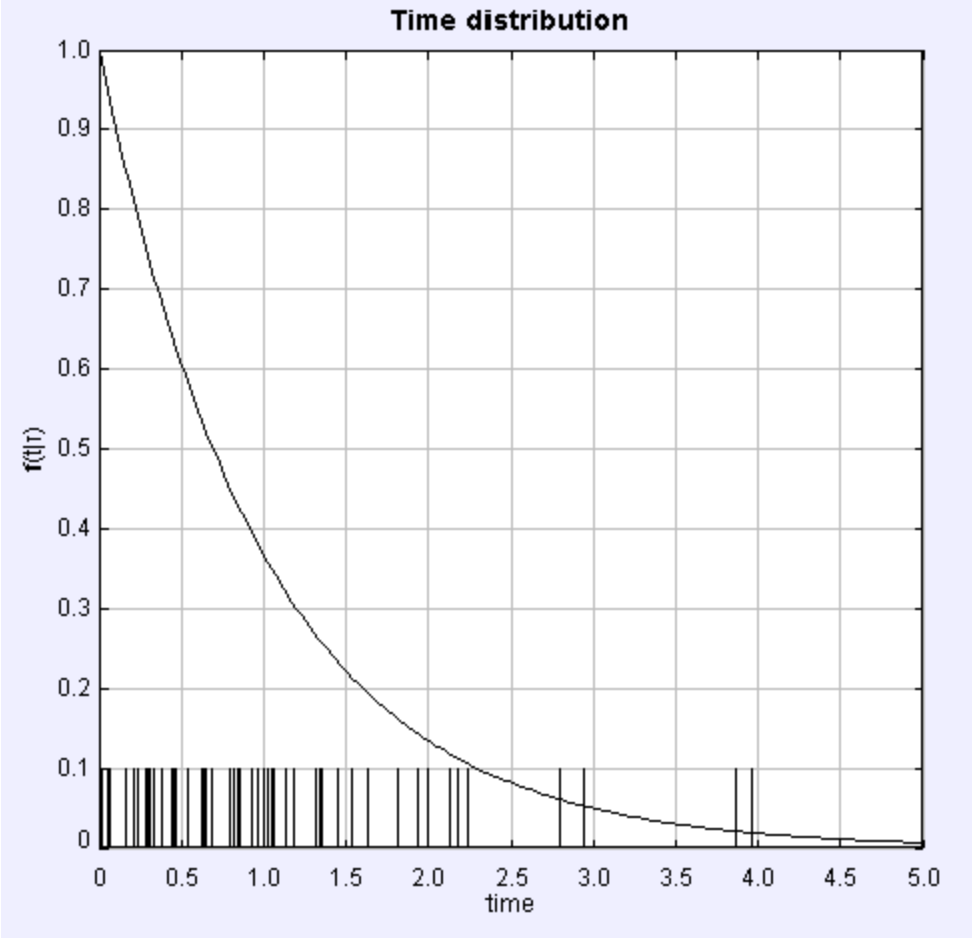
$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t}$$

Example of exponential sample:

- Shown below are 50 random numbers generated according to the pdf with $\tau = 1$.

  - ML estimate for this sample is

$$\hat{\tau} = 1.077$$



ML estimators for a Gaussian distribution

- The log-likelihood for measurements modelled by a Gaussian distribution with mean μ and variance σ^2 is:

$$\ln L(\mu,\sigma^2) = \sum_{i=1}^{n}\left(\ln\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\ln\frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

  - The maximum likelihood estimators are found as usual:

$$\left.\frac{\partial \ln L}{\partial \mu}\right|_{\mu=\hat{\mu}} = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\left.\frac{\partial \ln L}{\partial \sigma^2}\right|_{\sigma^2=\hat{\sigma}^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2 \quad \text{(biased)}$$

Maximum likelihood estimators in general

- For more complex models of experiments, it is not possible to evaluate the maximum of the likelihood function analytically
  - in this case, use numerical methods to determine the maximum likelihood estimates:
    - most optimization methods find the minimum, not maximum
    - for large data sets the likelihood may fall outside the valid range of real numbers in your programming environment (eg. 10^{-2000})
    - $\rightarrow$ find the values of the parameters that minimize the negative of the log-likelihood function

Question: modified exponential

- Consider a model that describes measurements by a random variable with pdf,

$$f(x) = a^2 x e^{-ax} \qquad x \geq 0$$

  - What is the maximum likelihood estimator for the parameter a, from a sample of size n observations?

  - In deriving the form for the estimator, when the derivative of the log-likelihood function is taken, the term involving x alone appears to drop out. Does this mean that the form of the estimator is the same whether it is there or not?

Question: repeated rate measurements

- Suppose you measure the activity of a radioactive source by counting the number of decays in one hour.
- You repeat this m times. Your data sample is therefore

$$\mathbf{n} = (n_1, n_2, \ldots, n_m)$$

  - Work out the maximum likelihood estimator for the activity of the radioactive source from such a sample.

Question: uniform distribution

- Suppose the model for an experiment is that the outcomes follow a uniform distribution between 0 and d where d is unknown.

- What is the maximum likelihood estimator for d?

Variance of ML estimators

- The variance is a measure of the spread a random variable
  - the variance of an estimator is an indication of its accuracy
- For simple pdf's, the variance can be calculated analytically:
  - example, the variance in lifetime estimator (sample mean) is the variance of the sample mean

$$\hat{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n} T_i \;\Leftrightarrow\; \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

$$V\left[\hat{\mathrm{T}}\right] = \frac{V[T]}{n} = \frac{\tau^2}{n} \;\Rightarrow\; \hat{V}_{\hat{T}} = \frac{\hat{\tau}^2}{n}$$

Variance of ML estimators

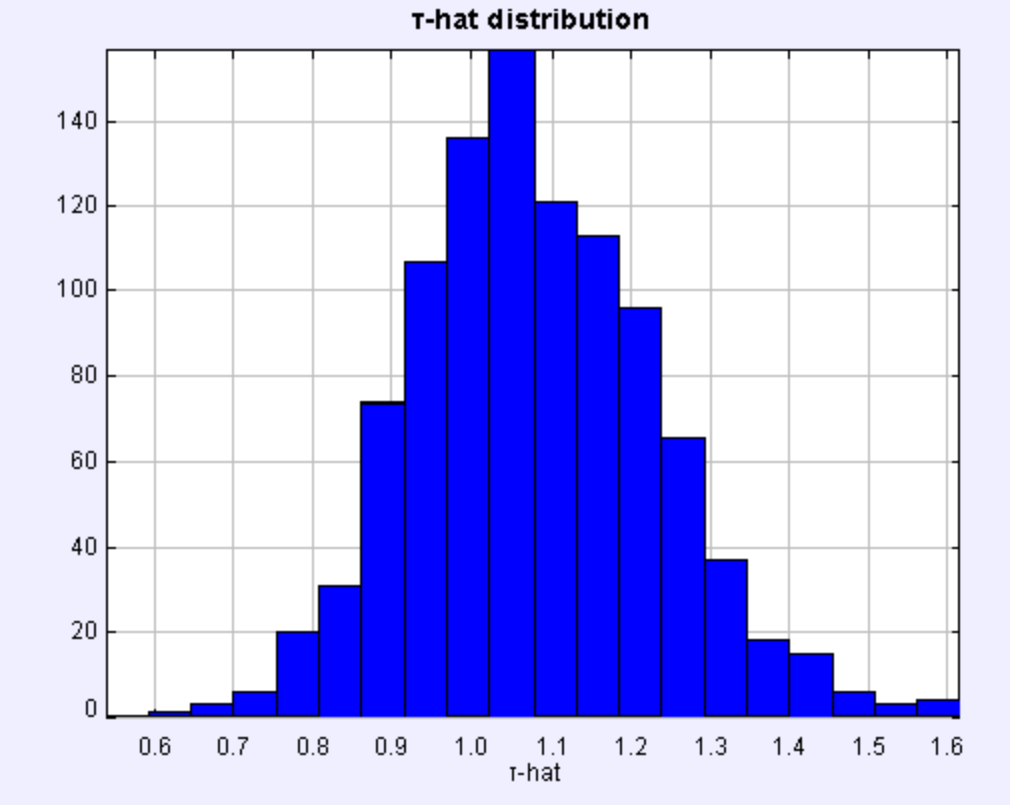
- For the hypothetical lifetime experiment with 50 measurements, the result could be reported as:

$$\hat{\tau} = 1.077 \pm 0.150$$

  - with the interpretation that the 1st number is the estimate and the 2nd number is the standard deviation of the estimator
    - Note: this is not the conventional use of this notation – to be discussed later

Variance of ML estimators: MC approach

- If it is too difficult to evaluate the variance of the ML estimators, another approach is to use Monte Carlo methods to simulate a large number of experiments, and estimate the variance from the sample of estimates
  - example: figure shows the estimates of 1000 repetitions of the experiment with 50 lifetime measurements with τ = 1.077



MC approach (cont.)

- Using this approach, m MC samples are used to estimate standard deviation of the lifetime estimator:

$$\hat{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} t_{ji} \qquad \bar{\hat{\tau}} = \frac{1}{m}\sum_{j=1}^{m} \hat{\tau}_j$$

$$\hat{\sigma}^2_{\hat{\tau}} = \frac{m}{m-1}\left(\overline{\hat{\tau}^2} - \bar{\hat{\tau}}^2\right)$$

  - the estimated standard deviation is 0.15
    - Agrees with the analytic result
- This approach cannot always be used: for complex experiments, large number of repetitions of the experiment could require too much computer time

The MC approach and the ensemble

- Repetition of MC experiments under the same conditions is possible, since it is a model
  - the meaning of "under the same conditions", however, may not be so clear
  - examples:
    - lifetime measurements: same amount of time or same number of events?
    - measurements with variable resolution: same number of events with each resolution, or not?
      - For example, an experiment that has probability of 0.01 that an event is measured with 100 times better resolution. Suppose the sample contains 9 regular events and 1 high resolution event – how do you choose the ensemble of experiments?

Variance of ML estimators: curvature

- In the large sample limit the likelihood function tends to follow a Gaussian:

$$L(\theta) \cong L_{\max} \exp\left(-\frac{(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2} \right)$$

$$\ln L(\theta) \cong \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2}$$

$$\Rightarrow \hat{\sigma}_{\hat{\theta}}^2 \cong -\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right)^{-1} \Bigg|_{\theta=\hat{\theta}} \qquad \hat{V}_{ij}^{-1} \cong -\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

  - Numerical methods are used to find both the minimum of the negative log-likelihood function and it's second derivatives at the minimum

Variance of ML estimators: graphical method

- If the likelihood function is approximately Gaussian, then

$$L(\theta) \cong L_{\max} \exp\left(-\frac{(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2} \right)$$

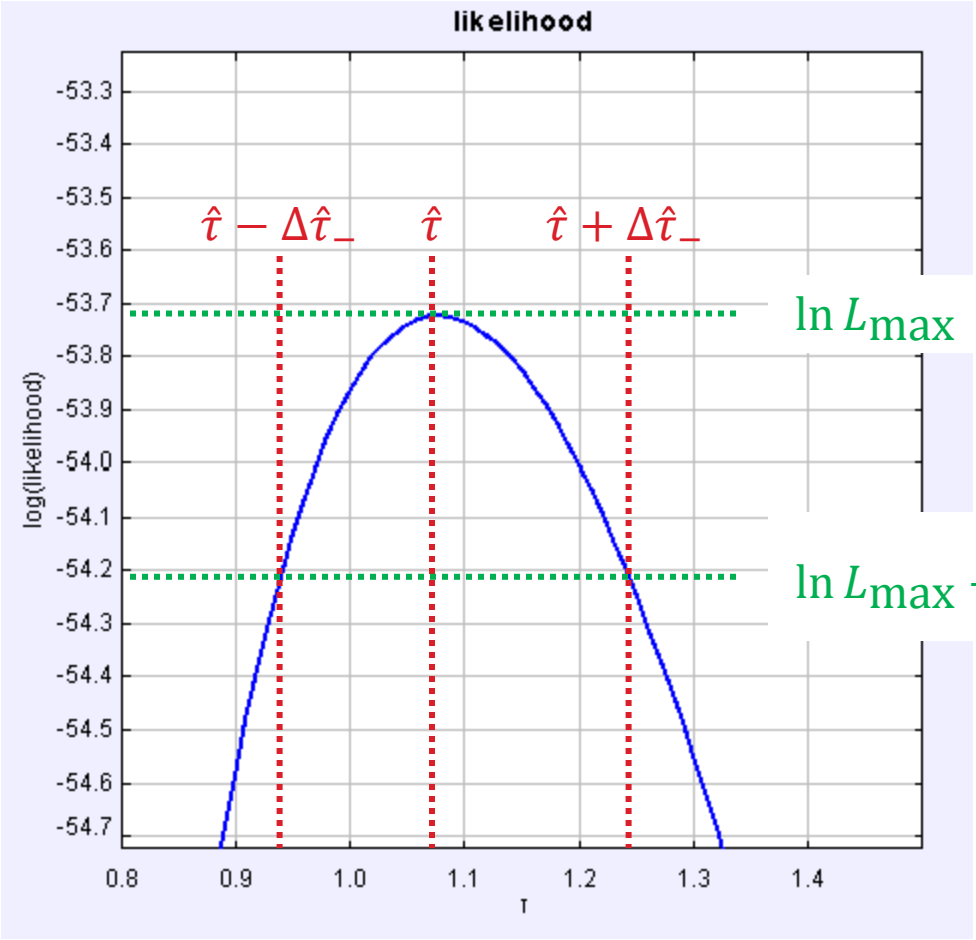
$$\ln L(\theta) \cong \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2}$$

$$\Rightarrow \ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \cong \ln L_{\max} - \frac{1}{2}$$

  - this approximation is very often used even when the likelihood function is far from Gaussian
    - not necessarily a good approximation!

Example of graphical method

- The lifetime experiment with 50 events:



$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \cong \ln L_{\max} - \frac{1}{2}$$

$$\Delta\hat{\tau}_- = 0.14$$

$$\Delta\hat{\tau}_+ = 0.16$$

$$\Delta\hat{\tau} = 0.15$$

$$\hat{\tau} = 1.08^{+0.16}_{-0.14}$$

$$\hat{\tau} = 1.08 \pm 0.15$$

Question: exponential

- Using the second derivative of the likelihood function, estimate the variance of the estimator for the lifetime.

Question: modified exponential

- Consider again the model that describes measurements by a random variable with pdf,

$$f(x) = a^2 x e^{-ax} \qquad x \geq 0$$

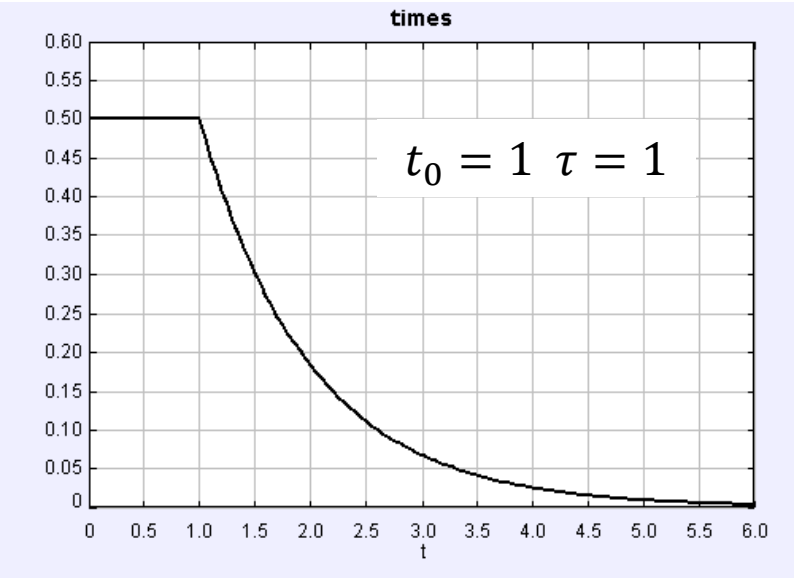
  - What is the variance of this estimator?
  - About how many observations are required to estimate a to a relative accuracy of 10%?

Example of ML with two parameters

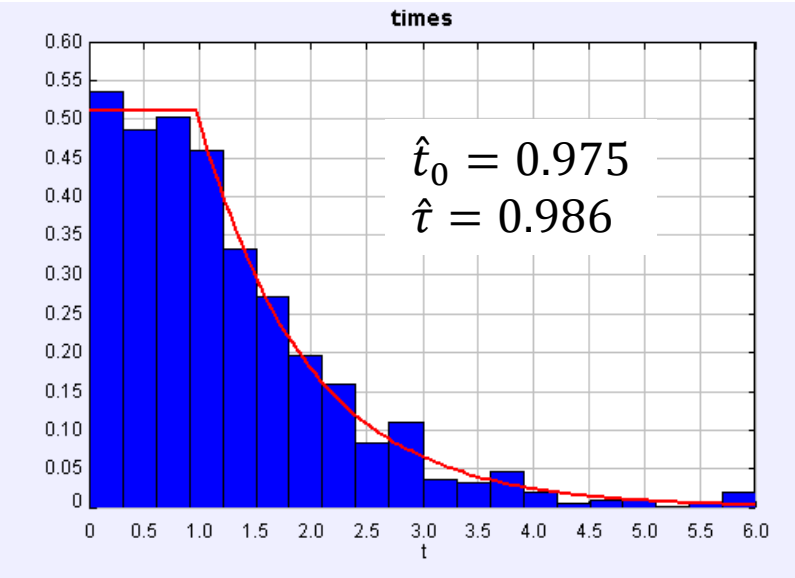
- Suppose that after being excited, a system emits a burst of light delayed by a random time. The probability distribution for the time delay is uniform for times between 0 and t_0, and thereafter falls off exponentially, with the decay constant, τ.

$$\text{PDF: } f(t|t_0,\tau) = \frac{1}{t_0+\tau} e^{-H(t-t_0)\frac{t-t_0}{\tau}}$$

ejs_fit2par.jar

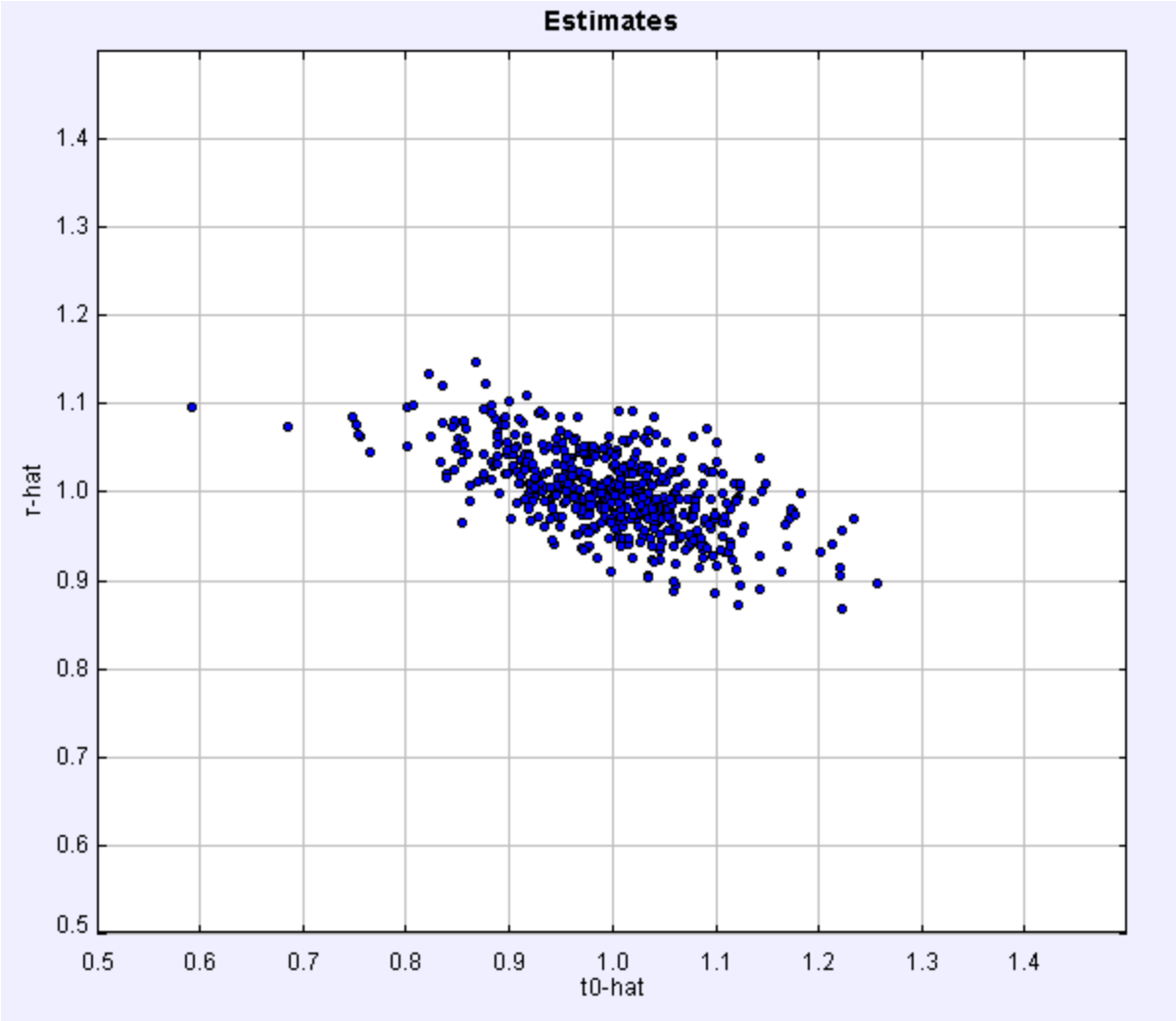


MC sample with 1000 events:



Example of ML with two parameters (cont.)

- Estimates from 500 similar experiments:
  - appear to be outcomes from 2D Gaussian



$$\bar{\hat{t}_0} = 0.996$$
$$\hat{\sigma}_{t_0} = 0.086$$
$$\bar{\hat{\tau}} = 1.001$$
$$\hat{\sigma}_{\tau} = 0.046$$
$$\hat{\rho} = -0.61$$

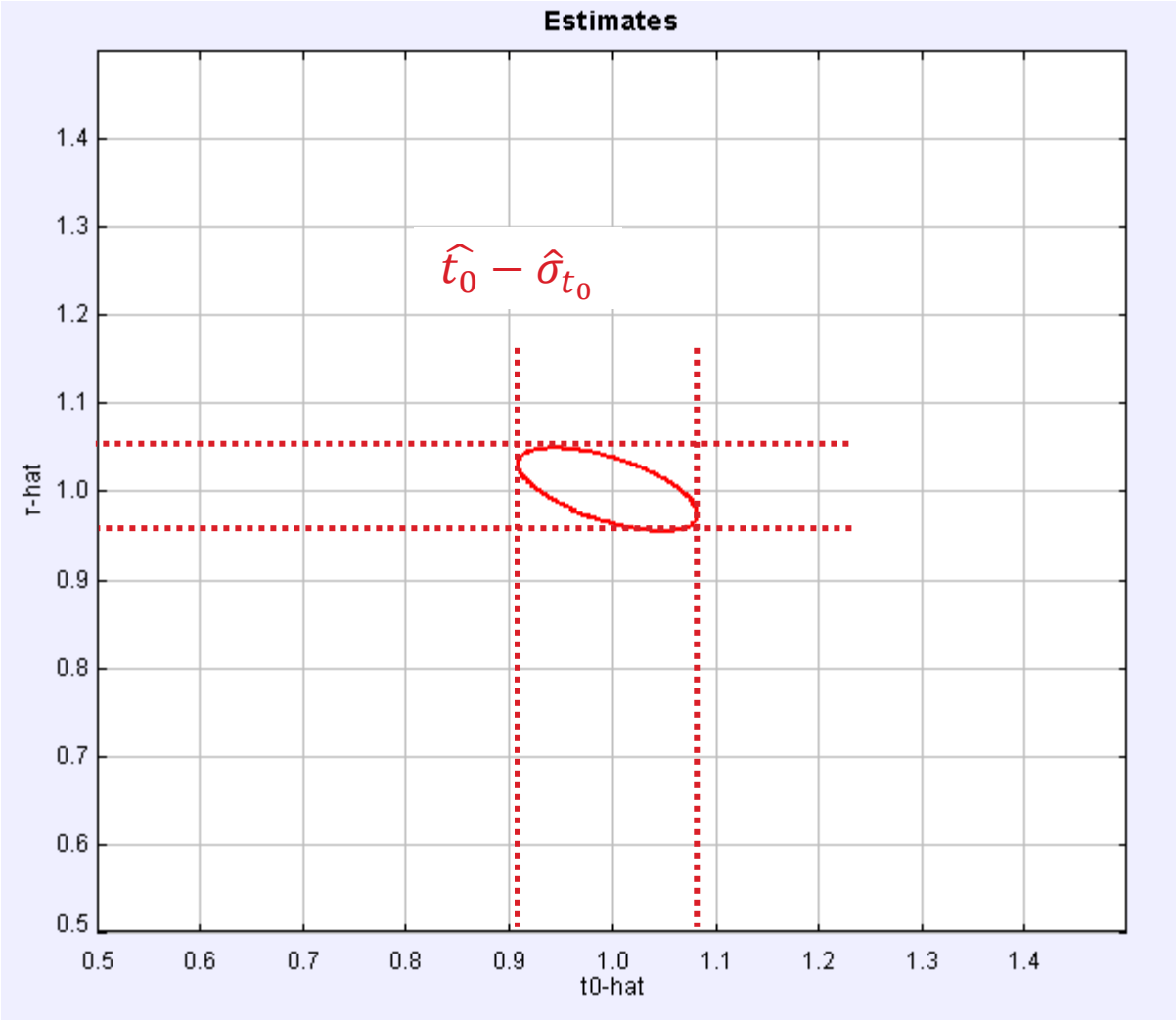
Graphical method with two parameters

- Sometimes the result is summarized by an ellipse, within which:

$$\ln L > \ln L_{\max} - \frac{1}{2}$$

  - example:

$$\widehat{t_0} = 0.996$$
$$\hat{\sigma}_{t_0} = 0.086$$
$$\hat{\tau} = 1.001$$
$$\hat{\sigma}_{\tau} = 0.046$$
$$\hat{\rho} = -0.61$$



Extended maximum likelihood

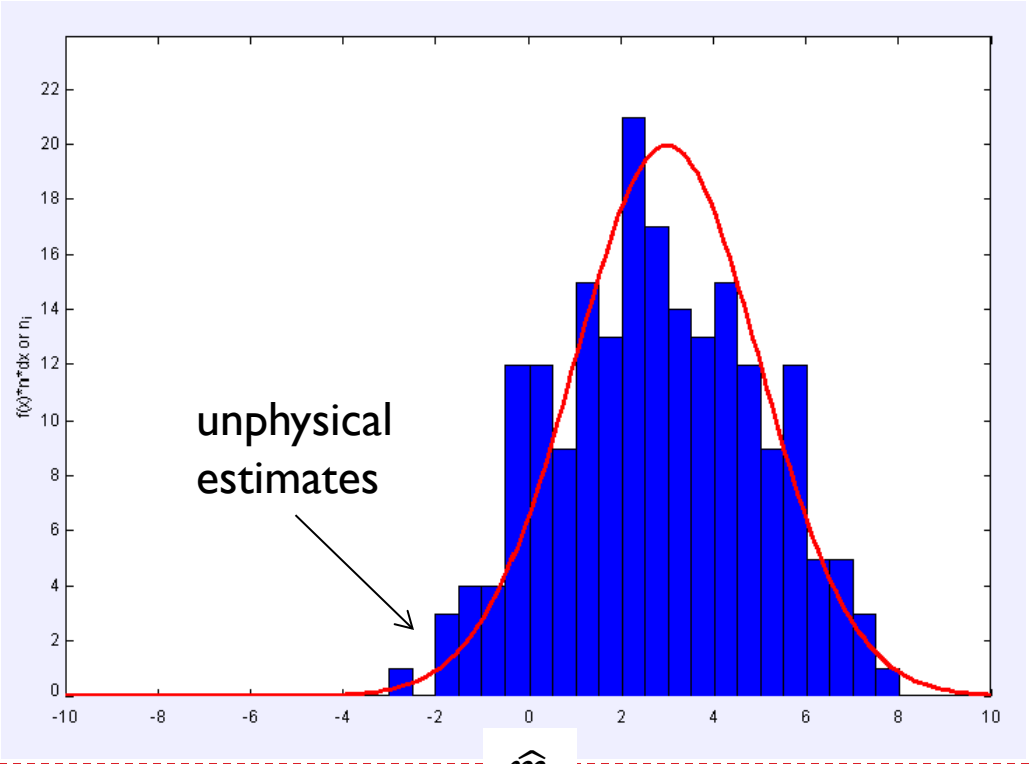
- This method is used in cases where model parameters define both
  - the distribution of events and
  - the event rate
- If the probability for an event to occur is constant in time, the number of events observed in a fixed time interval follows a Poisson distribution
  - In this case, the likelihood function is given by:

$$L(\nu, \theta) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i, \theta)$$

  - If $\nu = \nu(\theta)$ the variance of $\widehat{\Theta}$ will be reduced by using the extended likelihood function

Unphysical estimates

- Sometimes the ML estimates are unphysical
  - Example: mass estimate is negative
- It is important to report all estimates, even those that are unphysical, so that an average of experiments would not be biased
  - Example: distribution of estimates from 200 repetitions of an experiment



Maximum Likelihood and Bayesian Methods

- The maximum likelihood method is an ad hoc approach that has good properties in frequentist statistics
- In the Bayesian approach, it arises directly from Bayes theorem:

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

$$P(\theta \mid \vec{x}) = \frac{P(\vec{x} < \vec{X} < \vec{x} + d\vec{x} \mid \theta)}{P(\vec{x} < \vec{X} < \vec{x} + d\vec{x})} P(\theta)$$

$$= \frac{f(\vec{x} \mid \theta)}{f(\vec{x})} P(\theta) \quad \text{where } f(\vec{x} \mid \theta) = L(\theta)$$

$$= \alpha\, L(\theta) P(\theta)$$

α is a normalizing factor, so $\int P(\theta \mid \vec{x}) d\theta = 1$

Maximum Likelihood and Bayesian Methods

- A Bayesian estimator for a parameter can be the maximum of the posterior probability density:

$$P(\theta \mid \vec{x}) = \alpha \, L(\theta) P(\theta)$$

  - If the prior probability density is uniform (or approximately uniform, over the range that the likelihood function is large), the Bayesian estimator is the maximum likelihood estimator
    - in the case that the data is "stronger" than the prior belief, the Bayesian and Frequentist approach are the same

Question: 2 parameter example

- Explain how to generate MC events according to the problem with pdf:

$$f(t|t_0,\tau) = \frac{1}{t_0+\tau} e^{-H(t-t_0)\frac{t-t_0}{\tau}}$$

- What is the likelihood function?
- Can the maximum likelihood estimates be determined analytically?

Question: Extended maximum likelihood

- At time $t = 0$, an accelerator produces a very large number, m, of a particular radioactive isotope within a detector. The detector records the time of all decays. After a time period T, a total of n decays were recorded, a tiny fraction of m.
- Find the maximum likelihood estimators for the isotope lifetime using:
  - Only the recorded times: $(t_1, t_2, \ldots, t_n)$
  - Only the number of isotope decays observed: n
  - Both the recorded times and the number of decays observed
- Compare the variance of the estimators
  - Which has the smallest variance?

Method of Least Squares

D. Karlen / University of Victoria and TRIUMF

Method of Least Squares

- A special case of the maximum likelihood method
  - To be used when the model describes measurements as outcomes of Gaussian random variables
- The simplest model:
  - predicts the means of the Gaussian distributions
    - a function of the values of a controlled variable and unknown parameters (that are to be estimated)
  - standard deviations are the same for each random variable
  - random variables are independent
  - the controlled variable is not random – it is adjusted during the experiment to study its effect on the measurement
- For example: temperatures measured along the length of a bar
  - controlled variable: x_i the distance along the length of the bar
  - measurements: y_i is the temperature at each point
  - unknown parameter: thermal conductivity

Method of Least Squares

- If the model defines the expectation for each measurement as $E[Y_i] = \lambda(x_i, \theta)$ then the likelihood is given by,

$$L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(y_i - \lambda(x_i, \theta)\right)^2}{2\sigma_i^2} \right)$$

$$\ln L(\theta) = \text{const.} - \frac{1}{2} \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i, \theta)\right)^2}{\sigma_i^2}$$

  - So, the value of θ that maximizes the likelihood function, minimizes the chi-square function:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i, \theta)\right)^2}{\sigma_i^2}$$

    - "least squares" = least chi-square

Variance of the estimators

- The methods developed for the maximum likelihood method can be applied:
  - Curvature method:

$$\hat{\sigma}^2_{\hat{\theta}} \cong \left(\frac{1}{2} \frac{\partial^2 \chi^2(\theta)}{\partial \theta^2} \right)^{-1} \Bigg|_{\theta=\hat{\theta}} \qquad \hat{V}_{ij}^{-1} \cong \left(\frac{1}{2} \frac{\partial^2 \chi^2(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

  - Graphical method:

$$\chi^2(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \cong \chi^2_{\min} + 1$$

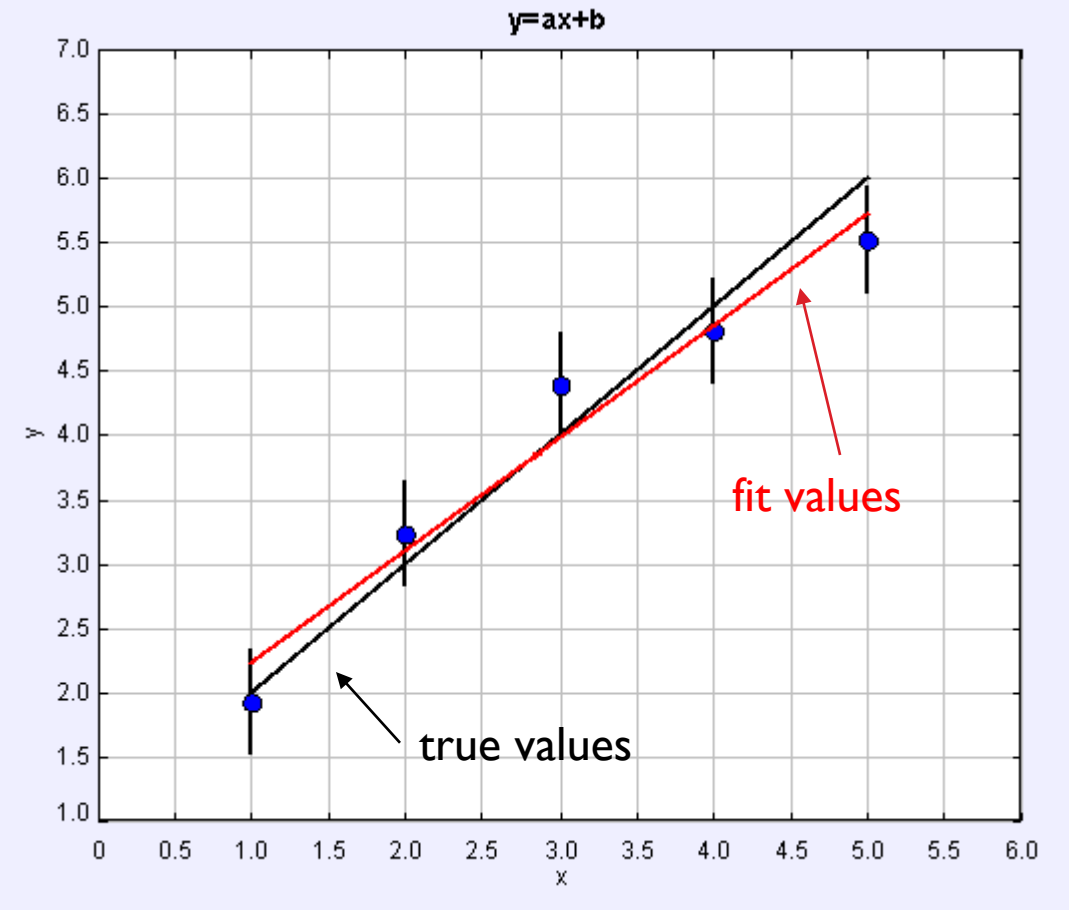
Example: Least squares fit to a line

- Assume independent random variables, Y_i , with the same standard deviation
- $\mathrm{E}[Y_i] = ax_i + b$

ejs_leastSquares.jar

$a = 1$
$b = 1$
$\sigma = 0.4$

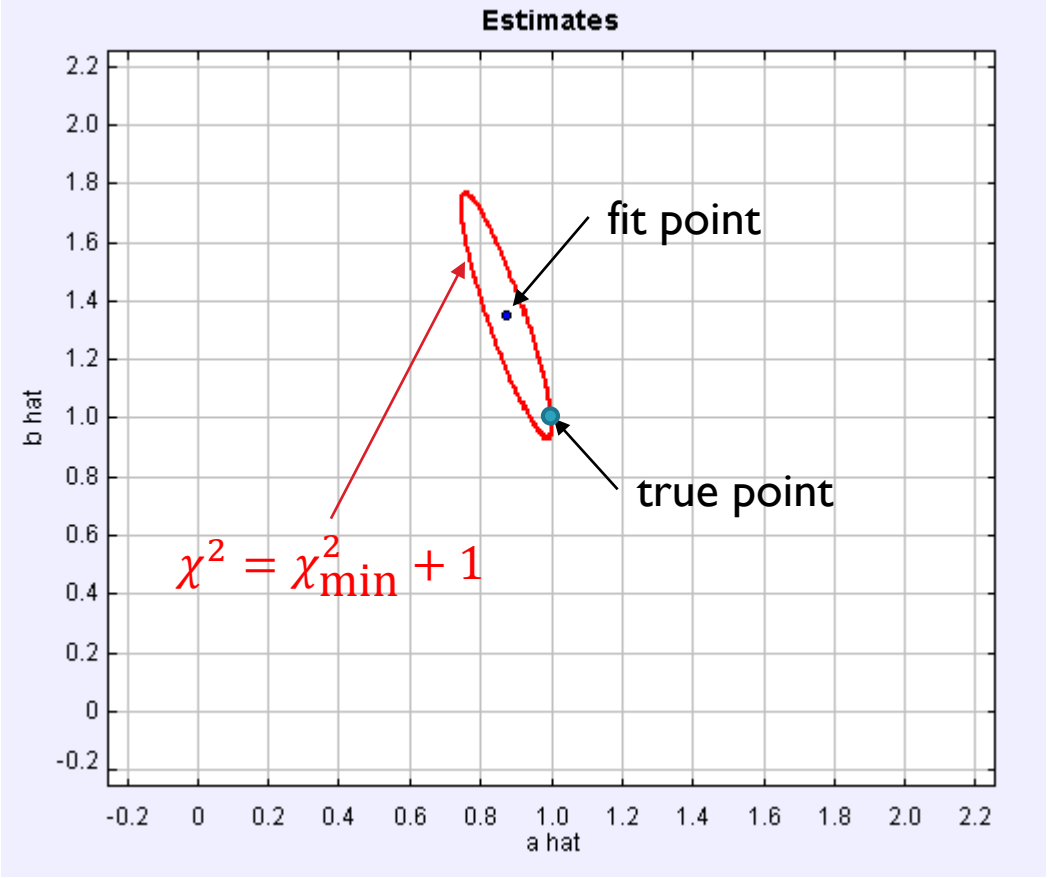
$\hat{a} = 0.87$
$\widehat{\sigma_a} = 0.13$
$\hat{b} = 1.35$
$\widehat{\sigma_b} = 0.42$
$\hat{\rho} = -0.91$



Example: Least squares fit to a line

- The covariance is illustrated by the ellipse:

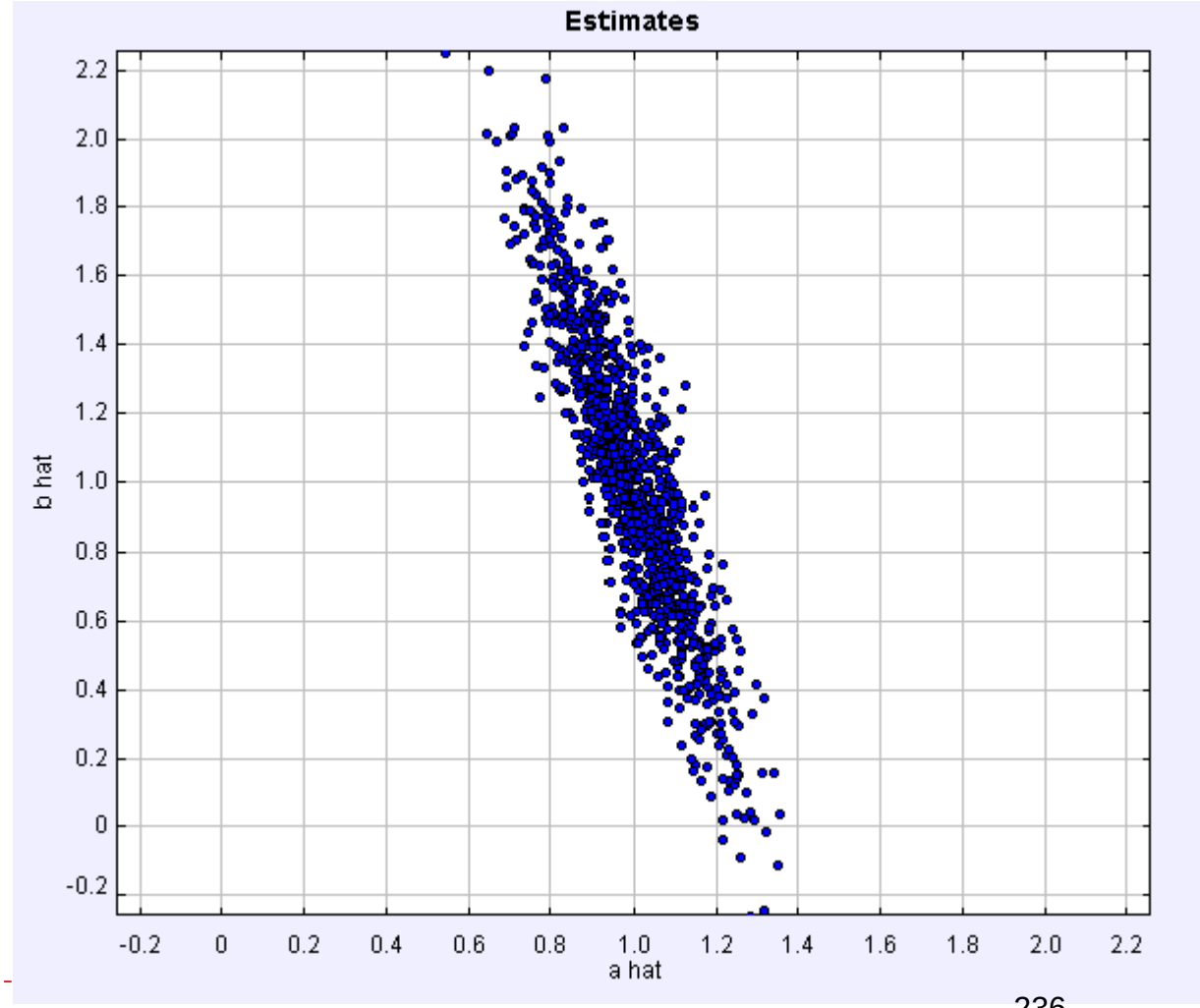
$\hat{a} = 0.87$
$\widehat{\sigma_a} = 0.13$
$\hat{b} = 1.35$
$\widehat{\sigma_b} = 0.42$
$\hat{\rho} = -0.91$



Example: Least squares fit to a line

- Repeated for 1000 experiments:
  - Estimates distributed according to a 2D Gaussian

$$\bar{\hat{a}} = 1.00$$
$$\sigma_{\hat{a}} = 0.12$$
$$\bar{\hat{b}} = 1.01$$
$$\sigma_{\hat{b}} = 0.42$$
$$\hat{\rho} = -0.89$$



Question: Least squares fit to a line

- Consider a model with $\sigma_i = \sigma$, and $\lambda = ax + b$, show that the least square estimates are given by:

$$\hat{a} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{V_{xy}}{V_x} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

- Derive the covariance matrix for the estimators

$$\chi^2 = \frac{1}{\sigma^2}\sum\left(y_i - (ax_i + b)\right)^2 = \frac{N}{\sigma^2}\left(\overline{y^2} + a^2\overline{x^2} + 2ab\bar{x} + b^2 - 2a\overline{xy} - 2b\bar{y}\right)$$

$$\frac{\partial\chi^2}{\partial a} = \frac{N}{\sigma^2}\left(2a\overline{x^2} + 2b\bar{x} - 2\overline{xy}\right)$$

$$\frac{\partial\chi^2}{\partial b} = \frac{N}{\sigma^2}\left(2a\bar{x} + 2b - 2\bar{y}\right) \Rightarrow a\bar{x} + \hat{b} - \bar{y} = 0 \Rightarrow \hat{b} = \bar{y} - \hat{a}\bar{x}$$

$$\hat{a}x^2 + \left(\bar{y} - \hat{a}\bar{x}\right)\bar{x} - \overline{xy} = 0 \Rightarrow \hat{a}\left(x^2 - \bar{x}^2\right) = \overline{xy} - \bar{x}\,\bar{y}$$

$$\hat{a} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{x^2 - \bar{x}^2}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Question: Least squares fit to a line

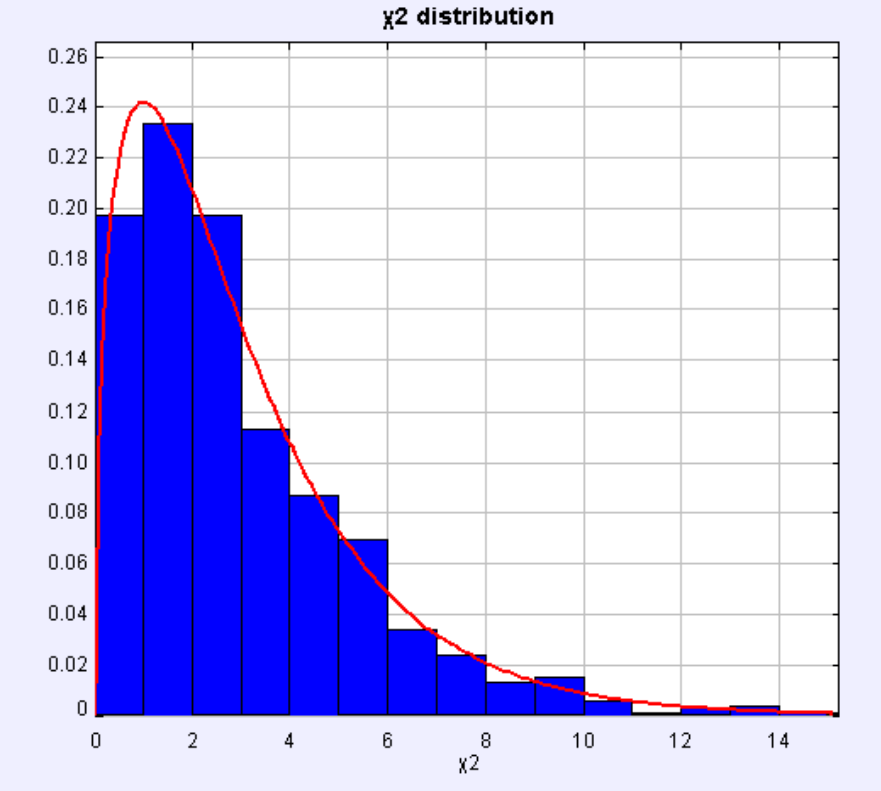
- Explain the features of the variances of the estimators:
  - It is not surprising that they depend on the variance of the random variables and on the number of random variables
  - Is it surprising that it does not depend on the means or outcomes of the random variables?
- Explain the interesting features of the correlation between the estimators:
  - Why is it negative?
  - Why does it not depend on the variance of the random variables?

Goodness of fit

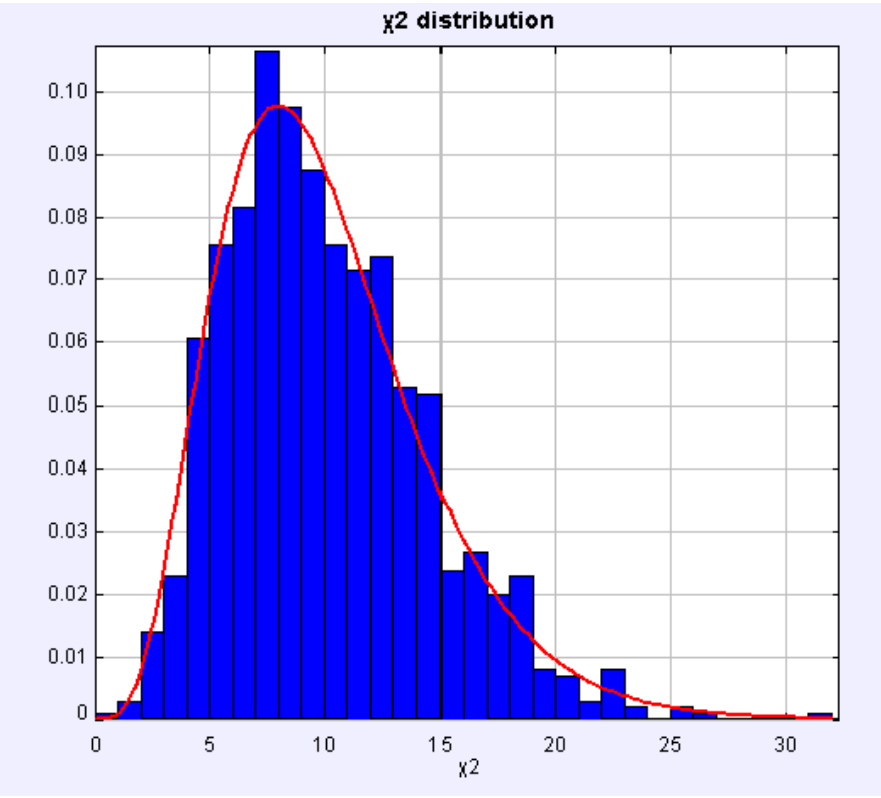
- The value of χ^2 at the minimum, reflects how well the data are compatible with the model:
  - assuming the model is correct, that there are n data points y_i, and that the parameterization λ is linear in the m parameters θ_j, the value of χ^2 at the minimum is an outcome of a random variable with pdf given by the χ^2 distribution with $n - m$ degrees of freedom
  - Derive the P-value: the probability of observing as large a χ^2 as seen or larger

Example: Goodness of linear fit

- χ^2 distribution for 1000 repetitions of the experiment:



5 points
nDOF = 3



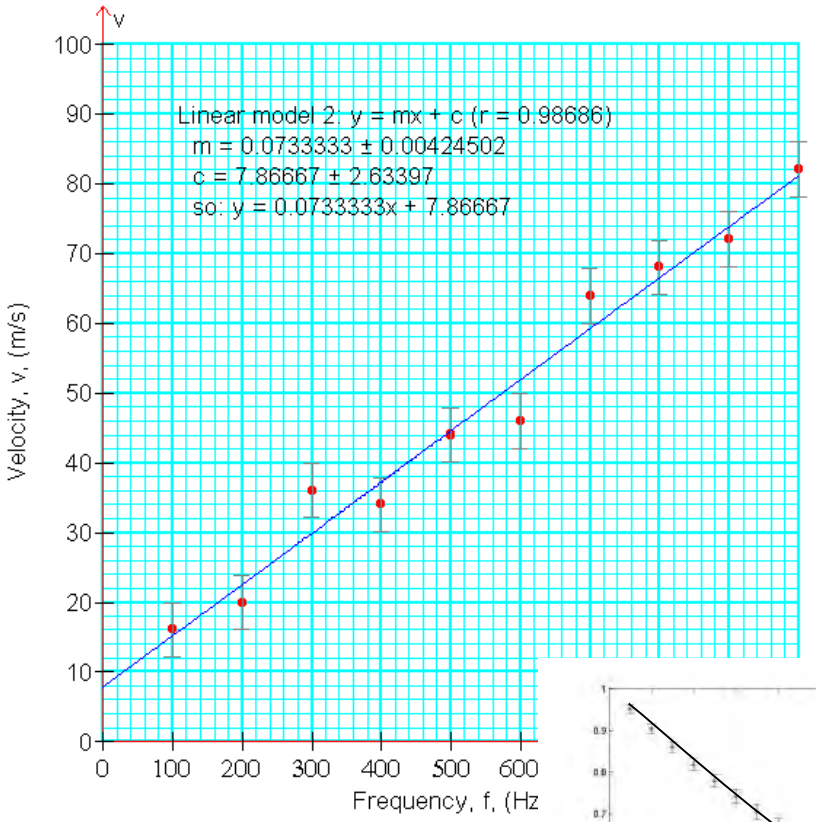
12 points
nDOF = 10

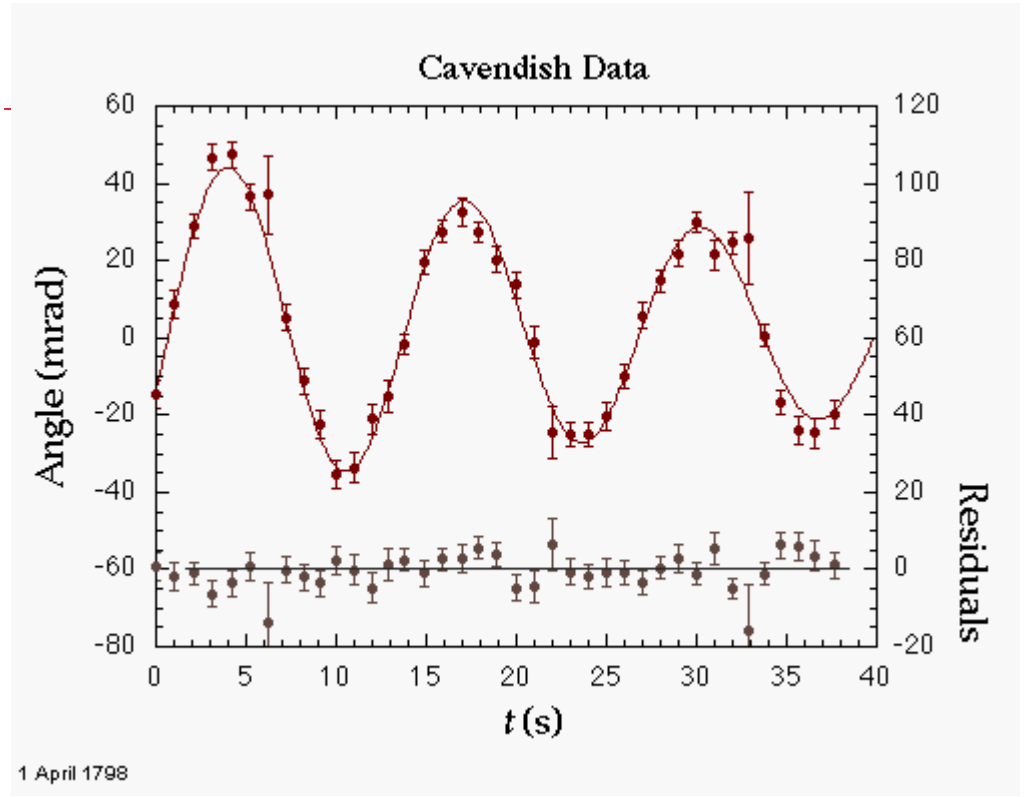
Visual goodness of fit

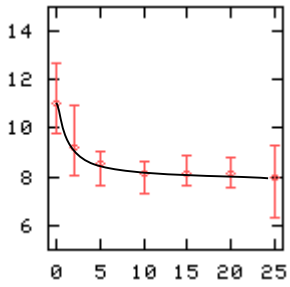
- For a data plot with many points, one can estimate the goodness of fit:
  - sum the number of standard deviations squared
  - check that about 68% of points are within 1 standard deviation of curve

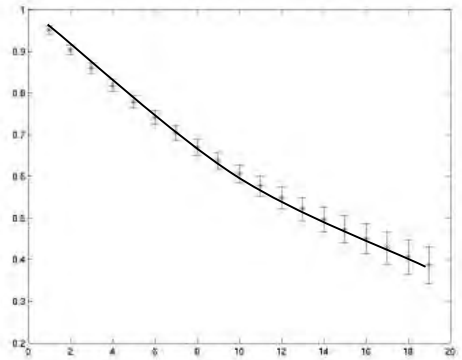
  - if P-value is too small: question the model
  - if P-value is too large: question the assigned uncertainties

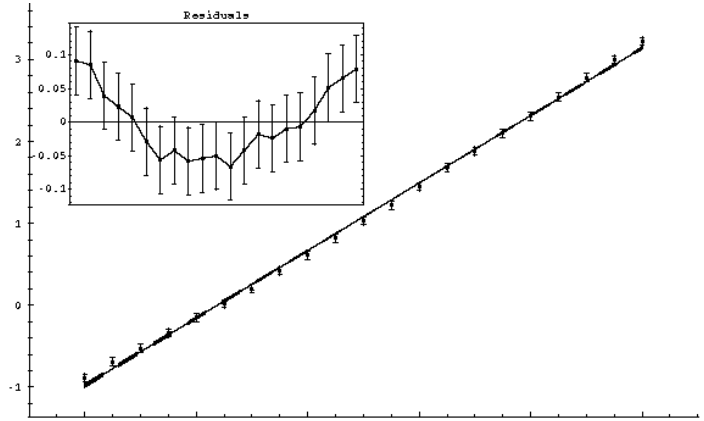
Misc examples





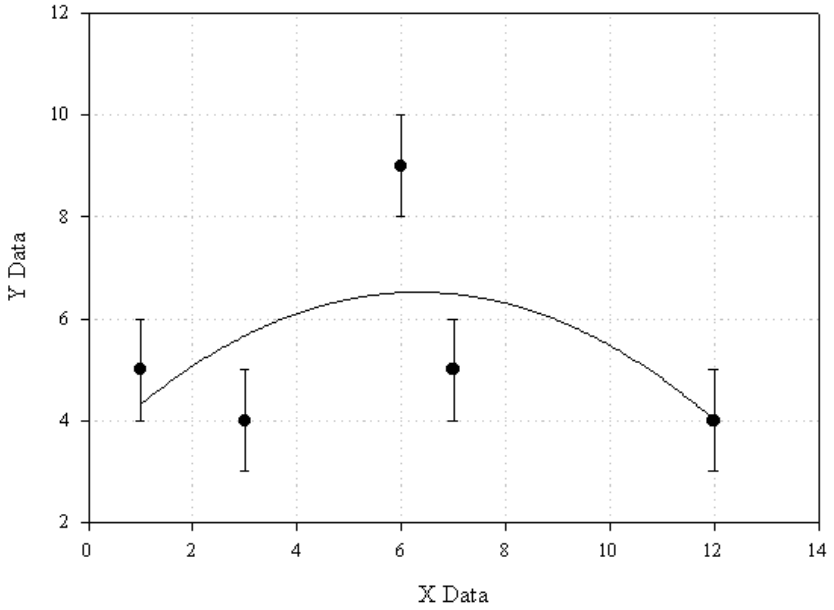






Question 1

- Figure shows data and model
  - χ^2 is about 11.1
  - What is the P-value for the goodness of fit?



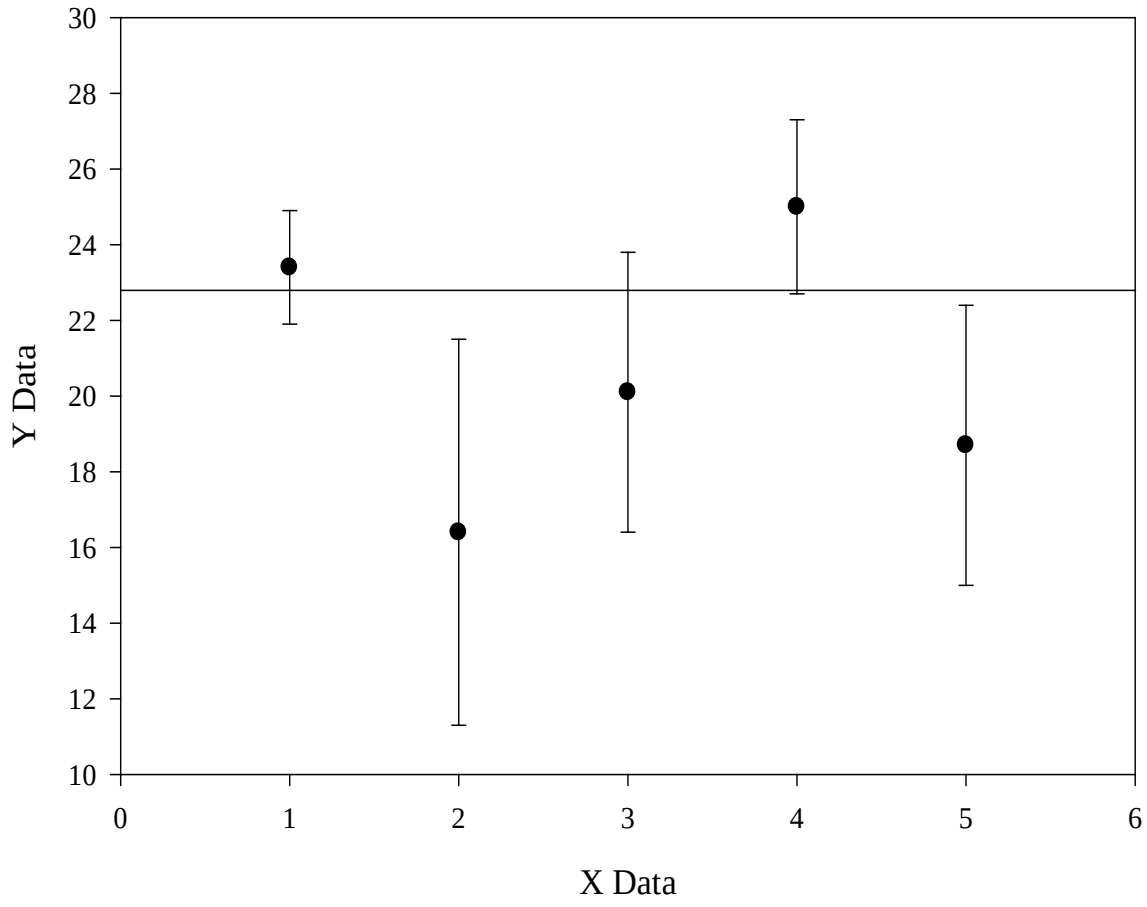
Question 2

- Consider the following data:

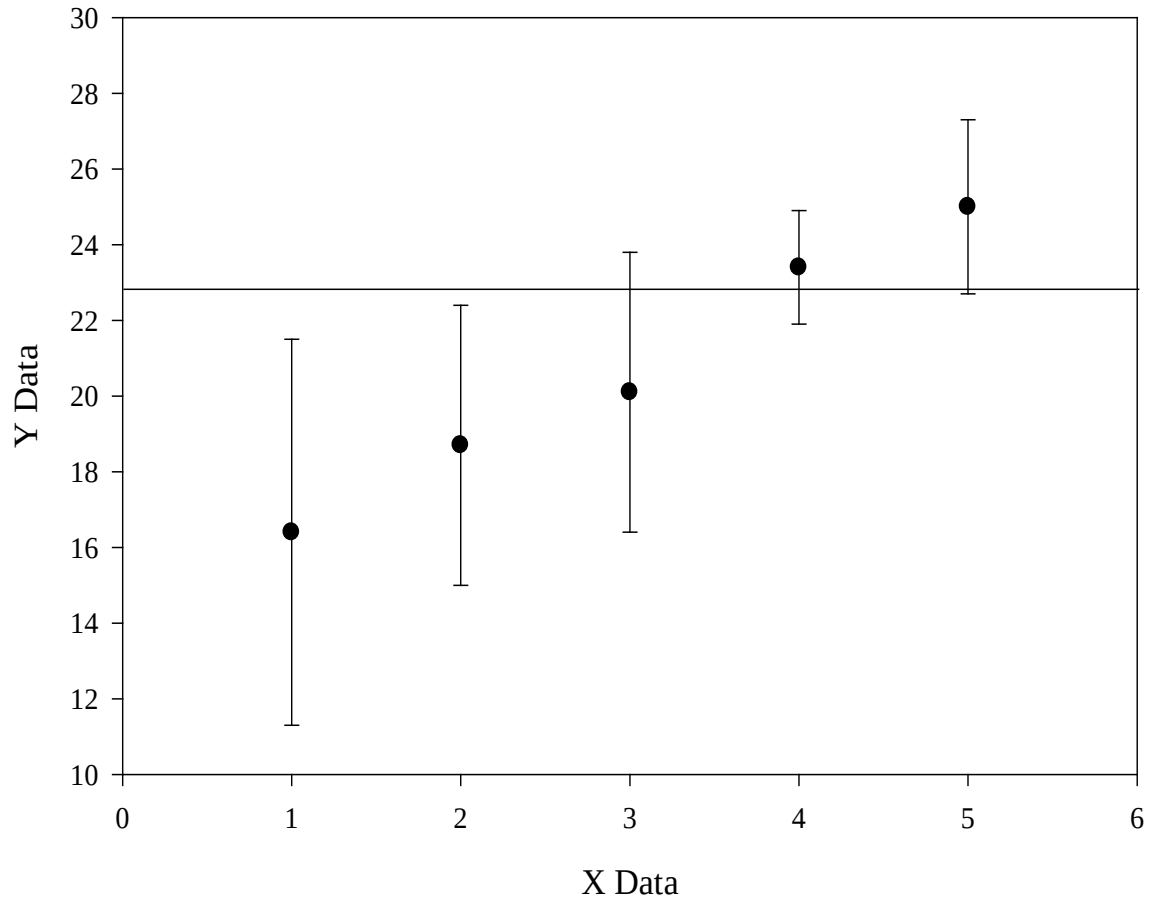
x	y	σ_y
1	23.4	1.5
2	16.4	5.1
3	20.1	3.7
4	25.0	2.3
5	18.7	3.7

  - if a zero-order LS fit is performed: $\chi^2 = 4.4$ for 4 d.o.f.
    - P-value = 0.35

Question 2



Question 2



Least squares fit to binned data

- Consider an experiment that measures a single quantity, x, modeled by a random variable, X, that follows the pdf, $f(x|\theta)$. Analyze the sample of n outcomes, $x_1 \ldots x_n$:
  - Tabulate the n outcomes into bins $i = 1 \ldots N$
  - Minimize the function: $$\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i, \theta)\right)^2}{\sigma_i^2}$$
    - y_i is the number of entries in bin i
    - $$\lambda(x_i, \theta) = n \int_{x_{i\,\min}}^{x_{i\,\max}} f(x|\theta)dx \approx n\ f(x_i|\theta)\,\Delta x$$
    - $\sigma_i^2 = \lambda(x_i, \theta)$ (since y_i are outcomes from a Poisson)
    - an approximation: least squares assumes σ is not a function of θ

Combining measurements with L.S.

- Suppose N experiments performed to estimate a parameter: $\hat{\theta}_i, \sigma_{\hat{\theta}_i}$
  - Use least squares to zeroth order polynomial to best estimate the parameter:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{\left(\hat{\theta}_i - \theta\right)^2}{\sigma^2_{\hat{\theta}_i}}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{N} \hat{\theta}_i / \sigma^2_{\hat{\theta}_i}}{\sum_{i=1}^{N} 1/\sigma^2_{\hat{\theta}_i}} \qquad \sigma^2_{\hat{\theta}} = \frac{1}{\sum_{i=1}^{N} 1/\sigma^2_{\hat{\theta}_i}}$$

  - well known formula for weighted average

Least squares fit to correlated data

- If the N measurements are modeled by a general N-dim Gaussian distribution (with non-zero correlation), use the full pdf:

$$f(\vec{x} \mid \vec{\mu}, \mathbf{V}) = (2\pi)^{-N/2} \left|\mathbf{V}\right|^{-1/2} \exp\left[-\tfrac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \mathbf{V}^{-1}\left(\vec{x}-\vec{\mu}\right)\right]$$

means

covariance

determinant

inverse

$$\Rightarrow \chi^2(\theta) = \left(\vec{x} - \vec{\mu}(\theta)\right)^T \mathbf{V}^{-1}\left(\vec{x} - \vec{\mu}(\theta)\right)$$

Errors and Confidence Intervals

D. Karlen / University of Victoria and TRIUMF

Statistical "errors" (aka "uncertainties")

- We have discussed frequentist methods to estimate one or more model parameters, using experimental data
  - the "unpredictable" elements of experiments are modeled by random variables
  - use the model to define an estimator (a random variable)
    - the parameter estimate is an outcome of the random variable
- To convey more information about the result of the experiment, one can report the result as follows:

$$m = 203 \pm 4 \text{ g}$$

  - where the second value is the standard deviation of the estimator (also known as the standard error)
    - NOTE: this is not the conventional definition

Error propagation

- Sometimes an estimated parameter needs to be transformed, for example:
  - change of units: g to kg
  - experiment measures m^2 but you want to know m
- The error of the transformed parameter is estimated by error propagation
  - use the first order Taylor expansion to evaluate how much the transformed parameter changes

$$y(x) \cong y(x_0) + (x - x_0)\left.\frac{\partial y}{\partial x}\right|_{x=x_0}$$

  - example: x = m^2 then $y(x) = \sqrt{x}$

Error propagation

$$y(x) - y(x_0) \cong (x - x_0)\left.\frac{\partial y}{\partial x}\right|_{x=x_0} = a(x - x_0), \quad a = \left.\frac{\partial y}{\partial x}\right|_{x=x_0}$$

- Consider the possible outcomes for x
  - described by a random variable of mean x_0 and variance σ_x^2
- If x is an outcome of a random variable X, y can be considered to be an outcome of a random variable Y:

$$Y - y_0 = a(X - x_0)$$

- The variance of Y is therefore

$$\sigma_Y^2 = E[(Y - y_0)^2] = E[a^2(X - x_0)^2] = a^2\sigma_X^2$$

  - example: x = m^2 =256±9 g^2 then y = ? ± ?

Combining errors

- Often measurements are combined to estimate the value of a model parameter
- Examples:
  - measurements that estimate the same parameter:
    - make an average (or weighted average)
  - measurements that estimate different parameters:
    - measure the acceleration of an particle in a known electric field: this determines q/m. In a second experiment, measure the mass of the particle. To estimate the charge of the particle, multiply these two estimates.
- The error of the combination depends on the errors of the individual measurements and whether the measurements are independent

Combining errors

- The covariance is used to approximate the variance of a function of more than one random variable
  - again use the first order Taylor expansion:

$$\text{Let } \vec{X} = (X_1, X_2, \ldots, X_n)$$

$$Y(\vec{X}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (X_i - \mu_i)$$

  - then:

$$E[Y(\vec{X})] \approx y(\vec{\mu}) \qquad \sigma_Y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

    - we only need to know the mean and covariance of *X* in order to estimate the mean and variance of *Y*

Examples

- Sum:

$$Y = X_1 + X_2 \;\Rightarrow\; \mu_Y = \mu_1 + \mu_2 \;,\; \sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$$

  - this is exact, since higher order Taylor series terms are all zero
  - for a sum, "errors are added in quadrature"

- Product:

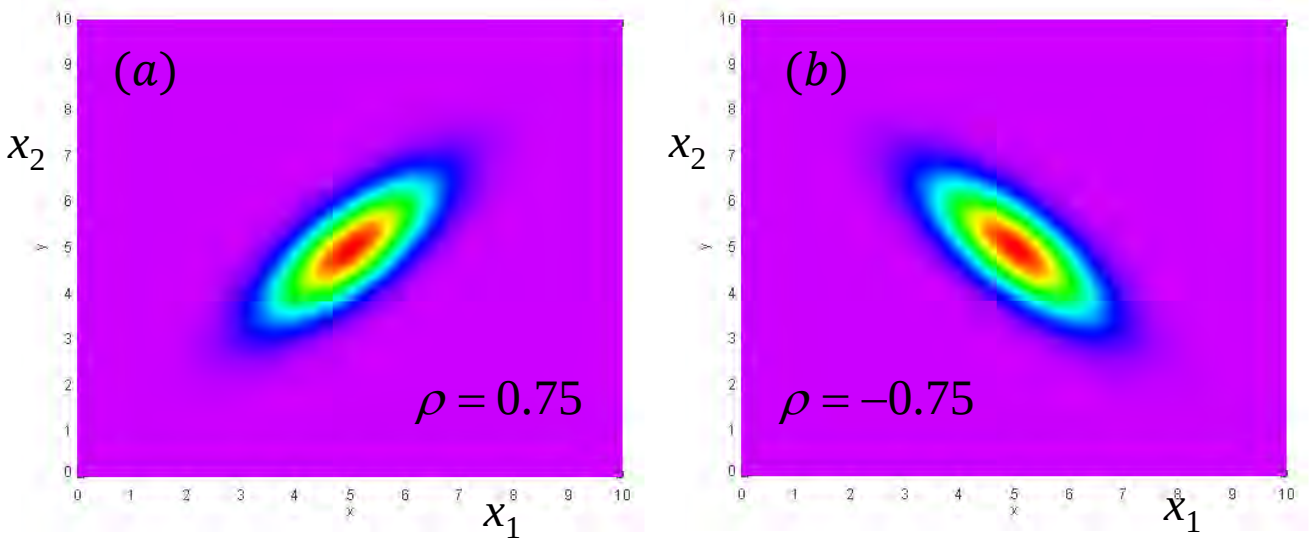
$$Y = X_1 X_2 \;\Rightarrow\; \mu_Y \approx \mu_1 \mu_2 \;,\; \frac{\sigma_Y^2}{\mu_Y^2} \approx \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} + 2\frac{V_{12}}{\mu_1 \mu_2}$$

  - for a product, "relative errors are added in quadrature"

Examples: Positive and negative correlation

- Consider the two cases below:

$\sigma_1 = \sigma_2 = 1$



$$Y = X_1 + X_2 \Rightarrow \quad \mu_Y^{(a)} = 10, V_Y^{(a)} = 3.5 \quad ; \quad \mu_Y^{(b)} = 10, V_Y^{(b)} = 0.5$$

$$Y = X_1 X_2 \Rightarrow \quad \mu_Y^{(a)} \approx 25, V_Y^{(a)} \approx 87.5 \quad ; \quad \mu_Y^{(b)} \approx 25, V_Y^{(b)} \approx 12.5$$

$$Y = X_1 / X_2 \Rightarrow \quad \mu_Y^{(a)} \approx 1, V_Y^{(a)} \approx 0.02 \quad ; \quad \mu_Y^{(b)} \approx 1, V_Y^{(b)} \approx 0.14$$

Question: Parameter transformations

- The mass-squared of a block is estimated to be:

$$m^2 = 256 \pm 9\ \mathrm{g}^2$$

  - How would you report the estimate for the mass?

$$m = ? \pm ?\ \mathrm{g}$$

- The cosine of a small angle is estimated to be:

$$\cos\theta = 0.95 \pm 0.01$$

  - How would you report the estimate for the angle alone?

- For both examples, check whether the transformed interval corresponds to the original interval.

Question: Combining errors

- Consider measurements of the resistance of a resistor undertaken by two people, one right after another.
  - Suppose repeated measurements give different results because:
    - variable quality of the contact between ohm-meter and resistor
    - slow variations of the temperature of the ohm-meter
  - To model this situation:, use two random variables to represent future measurements by the two people. Assume that the two measurements are done at the same temperature.

$$X = \mu + C_X + T \qquad E[C_X] = 0 \qquad E[T] = \mu_t$$

$$Y = \mu + C_Y + T \qquad E[C_Y] = 0$$

  - What is the variance of the average of the two random variables?

Answer: $V_C/2 + V_T$

Question: Positive and negative correlation

- Explain the interesting results from the examples of combing random variables with positive and negative correlation:
  - Why do the sum and products have smaller variance when the random variables are negatively correlated
  - Why is the situation reversed for the division

- To most accurately estimate the amount of evaporation from a beaker of water after heating, is it better to make the two measurements before and after with the same scale, or with different scales?
  - Explain in terms of correlation

Statistical vs Systematic Errors

- Frequentists treat statistical and systematic errors very differently
  - Statistical error: refers to those aspects that would cause identical repetitions of an experiment to yield different estimates of the parameter
    - eg. unpredictable (random) factors, quantum effects
  - Systematic error: refers to those aspects that would cause identical repetitions of an experiment to consistently yield a parameter estimate that differs from the true value
    - eg. calibration errors, incorrect model assumptions

- Often reported as: $m = 203 \pm 4\ (stat) \pm 3\ (sys)\ g$

Systematic errors

- Unlike statistical errors, the effect of systematic errors are not reduced by averaging several measurements
  - furthermore, the existence of systematic errors cannot be detected by seeing different outcomes of identical measurements
- To evaluate the magnitude of systematic errors, you need to consider the outcomes of 'altered' experiments or models
  - Suppose you measure the current in a circuit by measuring the voltage across a 100Ω resistor.
    - If the resistance of the resistor is not well known, this leads to a systematic error in the measurement of the current
    - In this case, you need to consider the outcomes of experiments that use different resistors

Systematic errors

- What range of 'alteration' is appropriate?
  - In the case of the resistor, the range of resistances should reflect our knowledge of its resistance
    - manufacture may specify the accuracy of their resistors (eg. 1%)
    - we may have measured the resistance with a meter that is known to be accurate to 1%
  - The situation could then be modelled by there being a pile of resistors, having a distribution of resistances with mean 100 Ω and standard deviation of 1 Ω. You choose one resistor, and make the measurement.
    - Hypothetically, you could repeat the experiment by selecting another resistor.
    - In this way, a systematic error is evaluated like a hypothetical statistical error.

Question: Geiger counter uncertainties

- Suppose you want to measure the activity of a radioactive source, using a Geiger counter that has an efficiency of 90±3%.
  - In 100 seconds the counter detects 49,235 counts. What is the best estimate of the activity and what are the statistical and systematic errors?

Intervals

- So far, our treatment of 'errors' has been simplified.
- A rigorous treatment of errors is made by considering intervals.
  - Interval = range of values
  - In Bayesian statistical analysis, one uses credible intervals
  - In frequentist statistical analysis, confidence intervals are used

- If someone uses the word "interval", instead of "error", they are probably being more rigorous!

Bayesian Intervals – Credible Intervals

- As a result of an experiment, the Bayesian updates his/her belief for the parameter:
  - posterior-belief $\propto$ likelihood $\times$ prior-belief

$$P(\theta|x) \propto L(\theta)\ \pi(\theta)$$

  - point estimate: the mode of the posterior-belief
  - a credible interval [a,b] is formed and contains a known amount of probability, for example

$$\int_a^b P(\theta \mid x)\, d\theta = 0.9$$

    - interpretation: "the degree of belief that parameter is within the fixed interval [a,b] is 90%"
    - note: many intervals can satisfy the relation

Bayesian systematic uncertainties

- Systematic uncertainty is handled naturally within the Bayesian framework
- The physics and detector model are described in terms of physics parameters, θ, and systematic parameters, η
  - Often the systematic parameters are not of interest, and are given the name "nuisance parameters"
- One simply marginalizes over the systematic parameters, to get the posterior degree of belief of the physics parameters alone:

$$P(\theta|x) = \int P(\theta, \eta|x)\, d\eta \propto \int P(x|\theta, \eta)\pi(\theta)\pi(\eta)\, d\eta$$

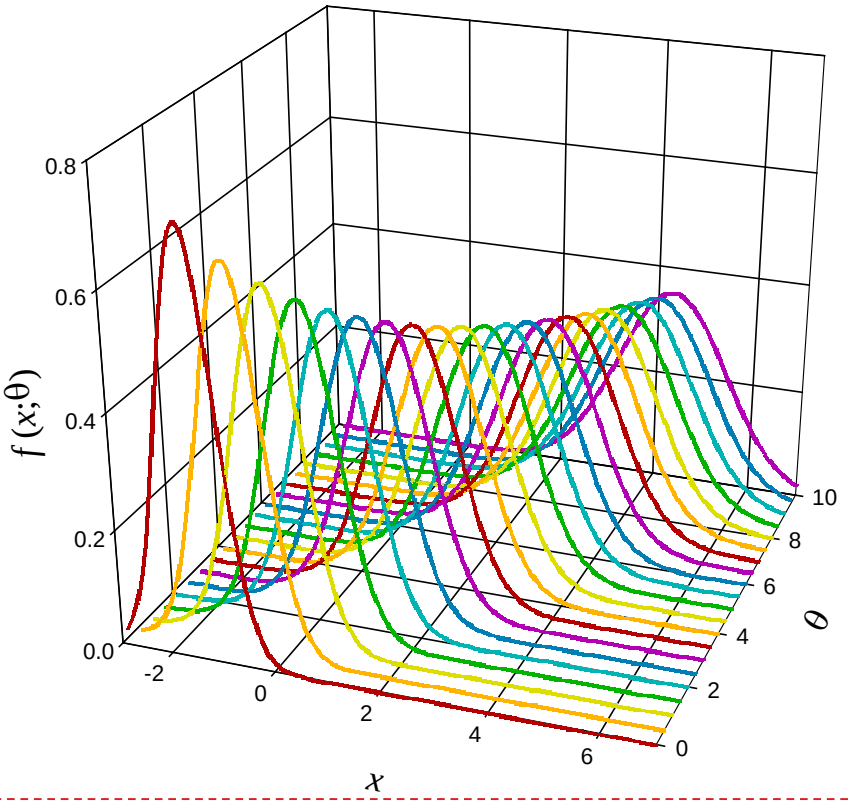
- Markov chain Monte Carlo is often used to derive the credible intervals of the marginalized posterior

Frequentist Intervals – Confidence Intervals

- Confidence Intervals are formed in such a way that they contain the true value(s) of the parameter(s) for a predetermined fraction of repeated experiments, say 90%
  - the boundaries of the intervals are modeled by random variables: $P(\Theta_a \leq \theta \leq \Theta_b) = 0.9$
  - interpretation: "90% of such intervals contain the true value"
    - they must satisfy this condition for any possible true value of the parameter (this property is known as "coverage")
  - in order to have correct coverage, the analysis procedure must be fixed before examining the data

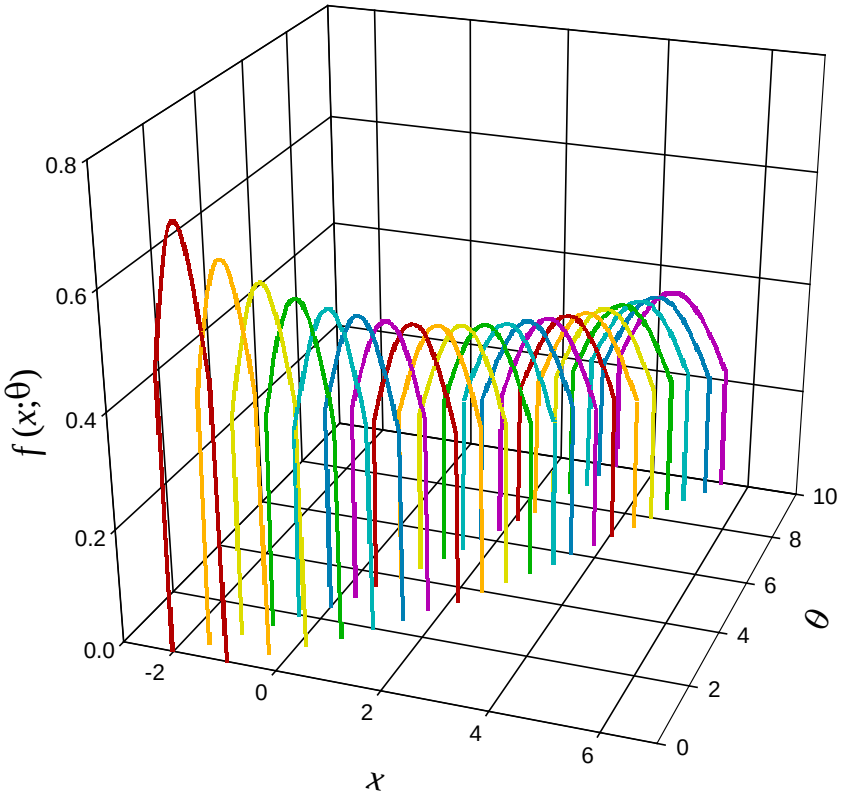
Constructing confidence intervals

- Neyman [1937] first defined the construction of such intervals:
  - consider measurements modeled by a random variable X, with pdf $f(x|\theta)$:



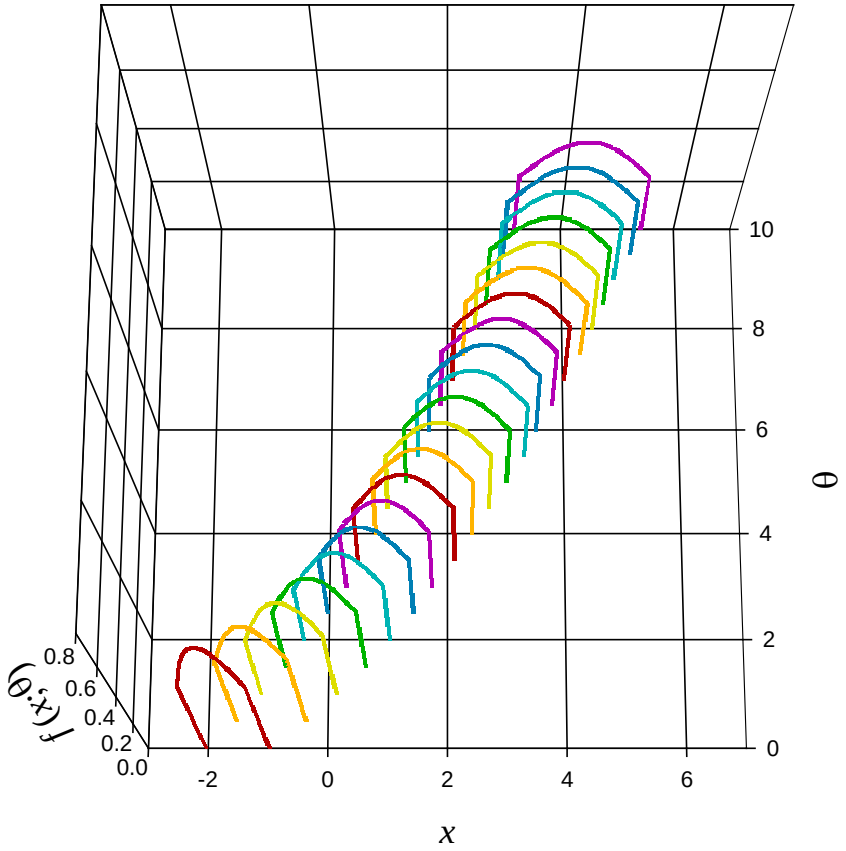
Constructing confidence intervals

- select a portion of the pdfs (with content α)
  - for example the 68% central region:



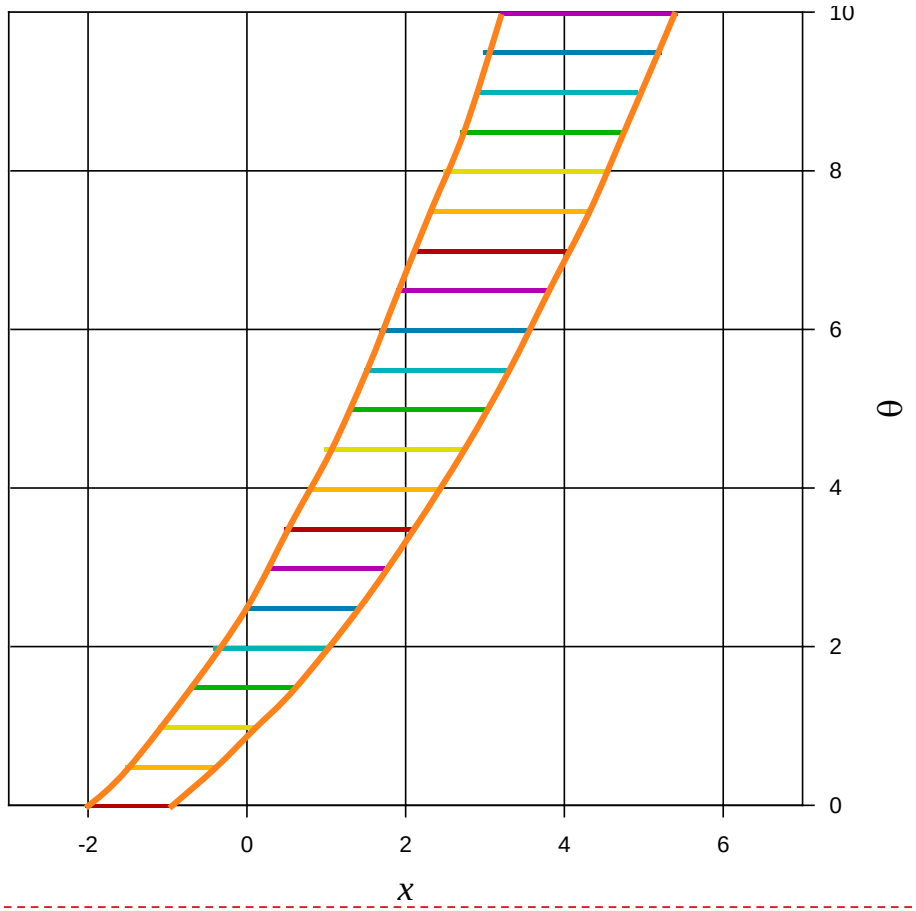
Constructing confidence intervals

- select a portion of the pdfs (with content α)
  - for example the 68% central region:



Constructing confidence intervals

- this gives the following confidence belt:



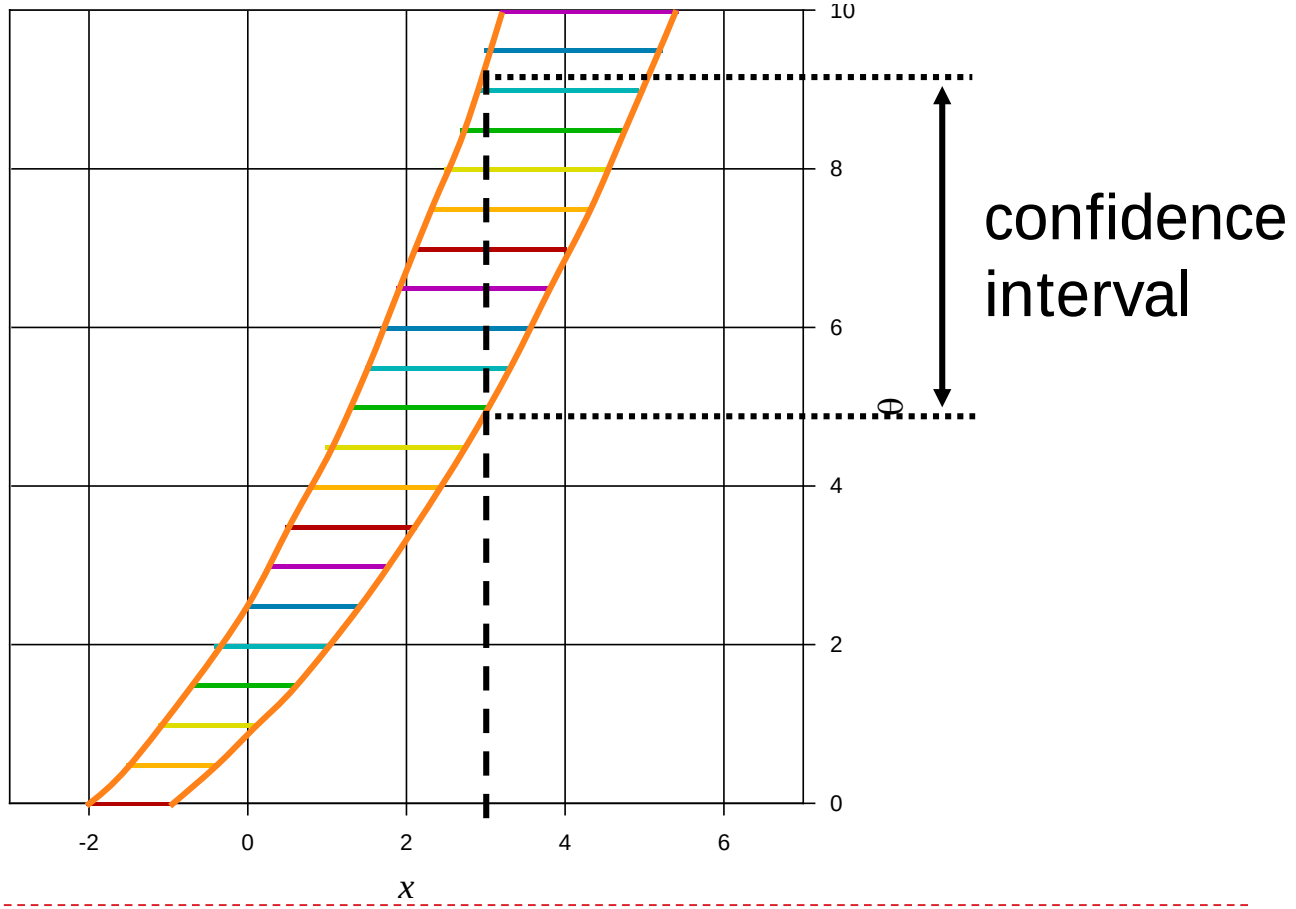
Constructing confidence intervals

- The (frequentist) probability for the random interval to contain the true parameter is α

α = fraction of experiments whose outcome is within the belt

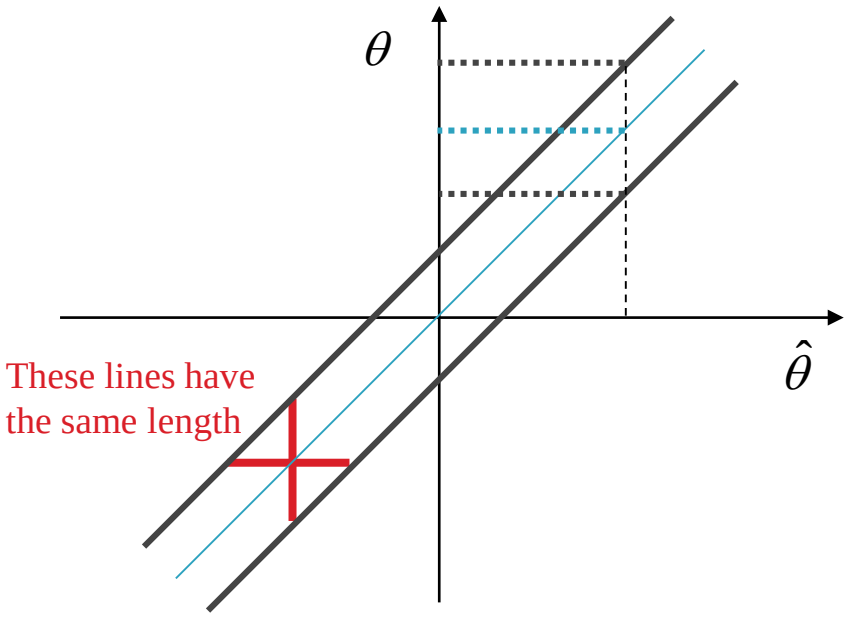
- only for those experiments, will the CI will contain the true value

so, the fraction of CI's that contain the true value = α



Confidence intervals: special cases

- The simplest case, frequently encountered, is when the estimator is unbiased and distributed as a Gaussian with variance independent of the true parameter:



• In this case, the 68.3% interval is the standard deviation of the estimator.

• This approach is often used even when the above conditions are not satisfied: an approximation

Confidence intervals: special cases

- For this special case, confidence intervals are easily calculated:
  - Central confidence intervals:

# σ	C.L.
1	68.3%
2	95.4%
3	99.7%

# σ	C.L.
1.645	90%
1.960	95%
2.576	99%

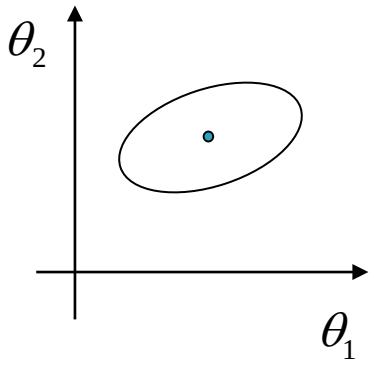
  - One sided confidence intervals:

# σ	C.L.
1	84.1%
2	97.7%
3	99.9%

# σ	C.L.
1.282	90%
1.645	95%
2.326	99%

Confidence intervals: special cases

- For this special case where there are more than one parameter, the confidence region can also be directly calculated by the covariance of the estimators:



Confidence region construction for n parameters:

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
C.L.	$\Delta \ln L$	$\Delta \ln L$	$\Delta \ln L$	$\Delta \ln L$
68.3%	0.5	1.15	1.77	2.36
90%	1.36	2.31	3.13	3.89
95%	1.92	3.00	3.91	4.75

Question: Mass of block

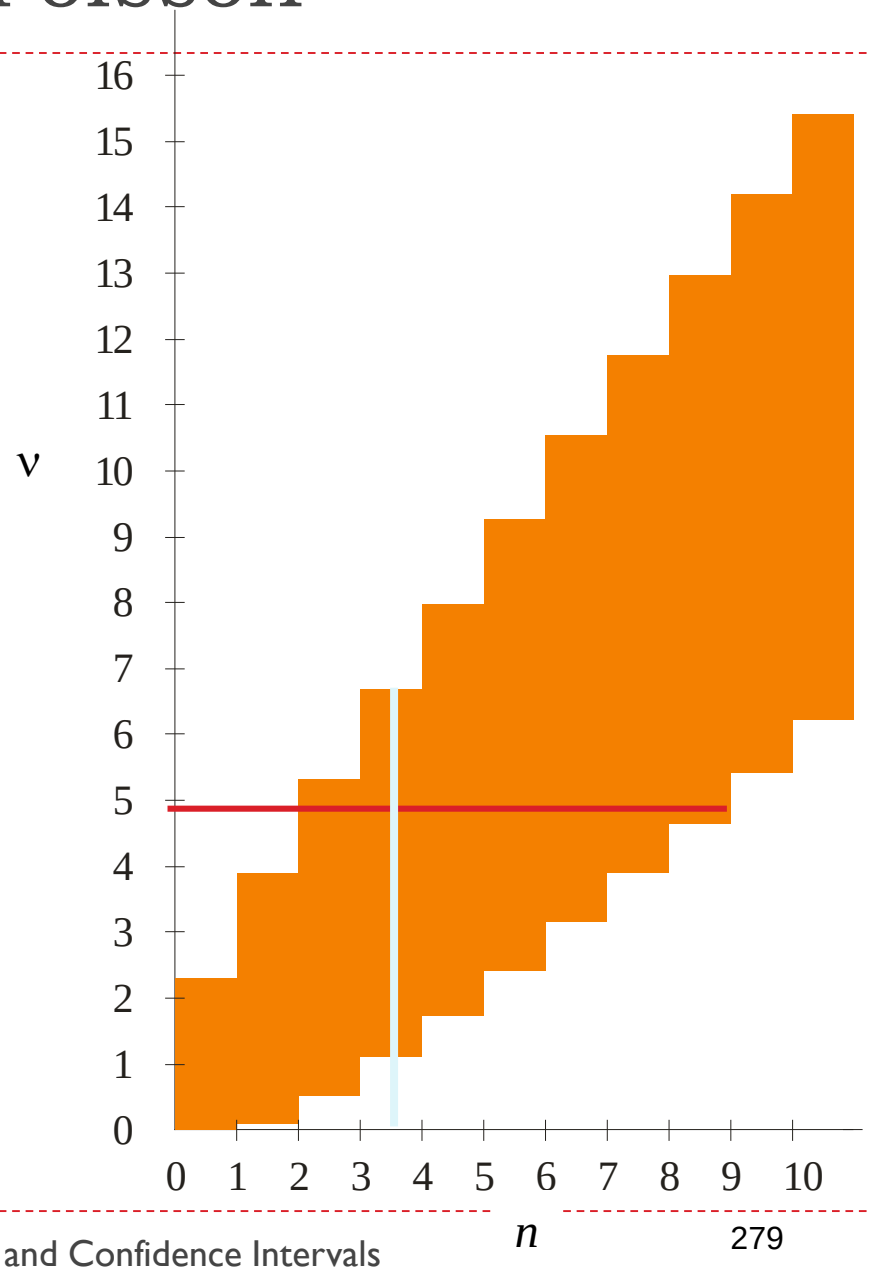
- An apparatus to measure the mass of a block is modeled by a random variable that is unbiased and has a standard deviation of 2 g. Suppose the mass of the block is measured 100 times, and the mean value of the measurements was 123.45 g.
  - What is the 95% central confidence interval for the estimate of the mass of the block?

Confidence intervals: Discrete RV

- For experiments with a discrete observable (like the Poisson distribution) the procedure has to be modified:
  - The portion of the pmf selected, for each true value of the parameter, must contain *at least* the stated confidence level.
    - Since observables are discrete, one cannot always form a simply connected set of them that have exactly a given probability
    - As a result, the confidence belt is "chunky" and the frequentist probability for the random interval to contain the true value is greater than the stated confidence interval (this is known as over-coverage)

Confidence belt for Poisson

- The central 80% confidence belt is indicated:
  - for any value of ν, at least 80% of the pmf is contained in the belt



for $\nu = 4.9$, the belt contains $n \in [2,8]$

for $n = 3$, the belt contains $\nu \in [1.1,6.7]$

Confidence intervals for Poisson

- For the Poisson distribution, the following table can be used to find the interval $[a,b]$:

$$\alpha = P(N \geq n_{\text{obs}} \mid a)$$
$$\beta = P(N \leq n_{\text{obs}} \mid b)$$

	lower limit a			upper limit b		
n obs	α=0.1	α=0.05	α=0.01	β=0.1	β=0.05	β=0.01
0	-	-	-	2.30	3.00	4.61
1	0.11	0.05	0.01	3.89	4.74	6.64
2	0.53	0.36	0.15	5.32	6.30	8.41
3	1.10	0.82	0.44	6.68	7.75	10.04
4	1.74	1.37	0.82	7.99	9.15	11.60
5	2.43	1.97	1.28	9.27	10.51	13.11
6	3.15	2.61	1.79	10.53	11.84	14.57
7	3.89	3.29	2.33	11.77	13.15	16.00
8	4.66	3.98	2.91	12.99	14.43	17.40
9	5.43	4.70	3.51	14.21	15.71	18.78
10	6.22	5.43	4.13	15.41	16.96	20.14

Question 1

- Consider an experimental analysis in which the estimator for the parameter θ is unbiased and follows a Gaussian pdf with variance

$$V = 0.25\,\theta^2$$

  - If the estimate is 3.0, what is the 68.3% central confidence interval for the parameter?

Question 2

- Consider an experiment that counts the number of occurrences in a Poisson process whose expectation value is 2.7. What fraction of measurements would report a central 80% confidence interval that contains the true value?

n	f(n|ν=2.7)
0	0.067
1	0.181
2	0.245
3	0.220
4	0.149
5	0.080
6	0.036
7	0.014
8	0.005

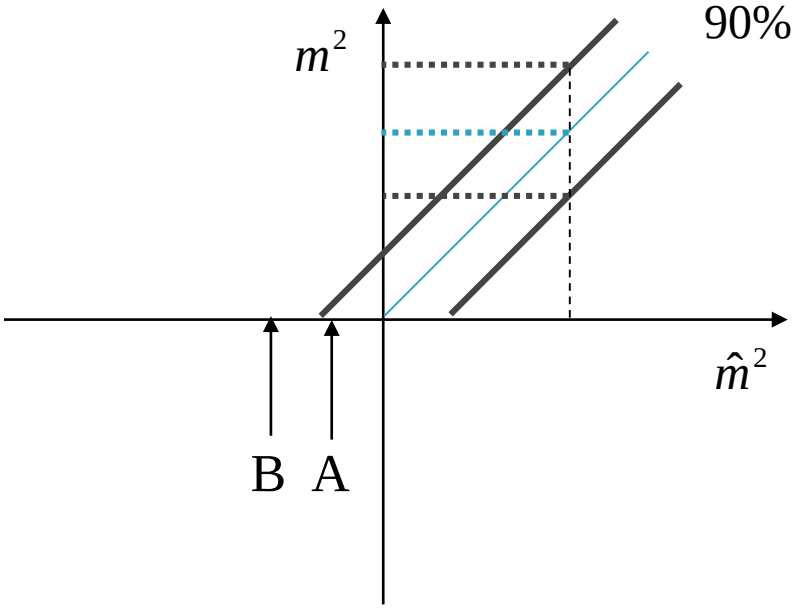
Confidence Intervals near a boundary

- Frontier experiments tend to search for new phenomena: effects that are at the edge of detectability (or beyond)
  - number of "signal" events observed above known background
  - mass of the electron neutrino
- In these cases there is a physical boundary
  - n_{signal} is not negative
  - m_ν is not negative
- The point estimate and interval can be in these non-physical regions:

Neutrino mass

- Use independent measurements of E and p:

$$\hat{m}^2 = E^2 - p^2$$

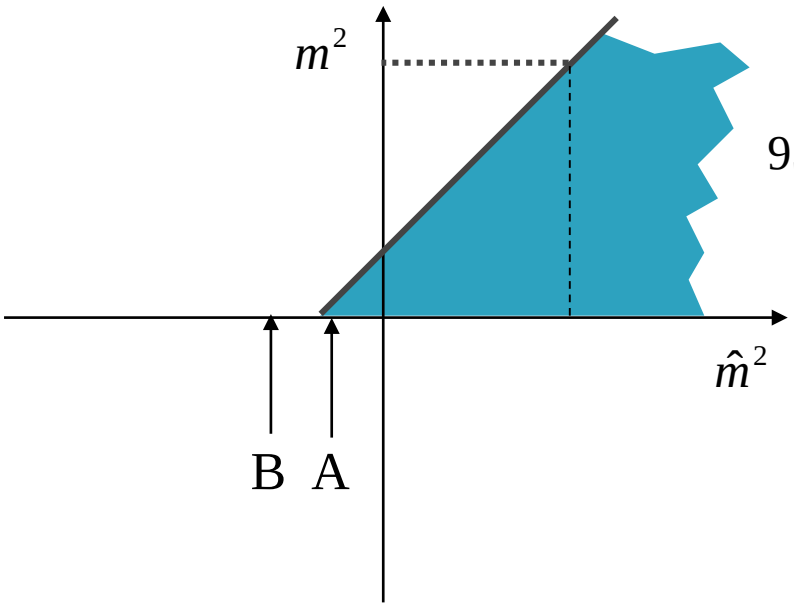


90% central confidence belt

What is confidence interval for experiment A? What about B?

Neutrino mass upper limit

- When it is not expected to measure the parameter, but to place a limit, a one sided belt can be used:



What is 95% CL upper limit for experiment A? What about B?

Special cases:

- Experiment A will have a very small upper limit
  - perhaps even smaller than the "competition" that had a better experiment!
- Experiment B will have an empty interval
  - The neutrino mass is in an empty interval at the 95% confidence level!
  - will happen for 5% of measurements if mass is zero
  - it is not valid to change CL to 99% so that upper limit is above zero
- One approach (even used by frequentist physicists!) is to use the Bayesian approach
  - use a uniform prior for $m_\nu > 0$

Summary

- Classical confidence intervals are designed so that a known fraction of them contain the true value
  - in order to ensure this behaviour, the choice of the confidence level (or any aspect of the analysis) must not depend on the data observed
  - some intervals are known not to contain the true value
    - need to be reported in any case, for averaging with other experiments
- Bayesian credible intervals
  - are never empty
  - rely on representing the prior degree of belief