Statistics

Point estimation

Let's say you collected some data. What you really care about is the probability distribution that *underlies* your data. But all you can do is sample a *finite* amount of data from the distribution. So how do you estimate a parameter (e.g. mean, variance) of the *underlying distribution* based on your *sampled data*? This process is known as calculating a **point estimate**. Although a point estimate might be an unfamiliar term, you are probably very used to calculating a point estimate for the mean or variance.

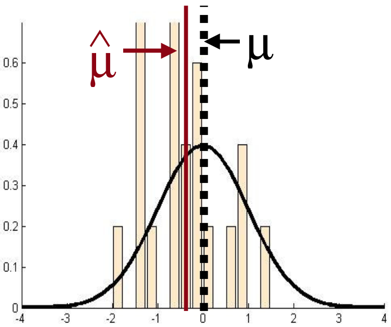
Example: Estimating the *mean* of the underlying normal distribution based on 20 data points ($y_1, y_2, \ldots y_{20}$).

μ = mean of the underlying distribution

$\hat{\mu}$ = our estimate of μ (the point estimate)

$$\hat{\mu} = \sum_i \frac{y_i}{20}$$

Notice that the estimated mean (usually) differs from the true mean:



Example: Estimating the *variance* of the underlying normal distribution based on 20 data points ($y_1, y_2, \ldots y_{20}$).

σ^2 = variance of the underlying distribution

$\hat{\sigma}^2$ = our estimate of σ^2 (the point estimate)

$$\hat{\sigma}^2 = \sum_i \frac{(y_i - \hat{\mu})^2}{20}$$

When plotting error bars, scientists often choose between plotting the **standard deviation** versus the **standard error of the mean**. What's the difference? The standard deviation is your estimate of the standard deviation of the actual underlying distribution. The standard error of the mean, on the other hand, is your estimate of the standard deviation of your *measurement* of the mean. Thus, the more points that go into your measurement of the mean, the smaller the standard error of the mean should be. On the other hand, the standard deviation should not depend on your sample size.

The first step of calculating either term is to calculate the point estimate of the variance:

$$\hat{\sigma}^2 = \sum_i \frac{(y_i - \hat{\mu})^2}{N}$$

N is the number of data points, y_i are the data points themselves, and $\hat{\mu}$ is the mean of the data points (i.e. point estimate of the mean).

In order to estimate the standard deviation (STD) of the *underlying distribution*, take the square root of the estimate of the variance:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

In order to estimate the standard deviation in *your estimate* of the mean, take the standard error of the mean (SEM):

$$\hat{\sigma}_{sem} = \frac{\hat{\sigma}}{\sqrt{N}}$$

The more data you have (the larger N), the smaller the SEM. That's because with more data, you have a better (less variable) estimate of the mean. The STD, on the other hand, does not change with the size of N.

Confidence intervals

Once you calculate a point estimate, how do you know how good an estimate of your underlying data it really is? If you calculated a point estimate of a mean, the SEM is one measure of how good your estimate is. But the **confidence interval** is a much more general and powerful approach.

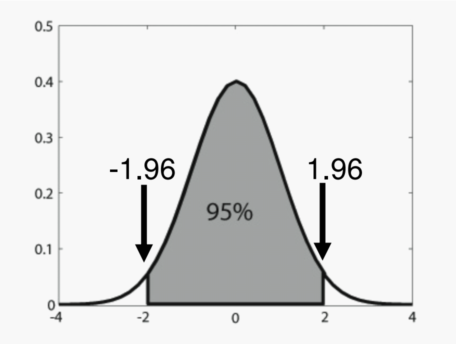
The **confidence interval** tells you the probability that the parameter of interest in the underlying distribution falls within some interval. More specifically, if you have taken a number of measurements of a parameter, the confidence interval will allow you to specify a range that with 95% probability contains the true value of the parameter.

Let's first consider the confidence interval of the mean when you have a distribution of *known* variance. To calculate the confidence interval, we take advantage of the Z distribution, which is simply a Normal (or Gaussian) distribution with mean of 0 and variance of 1.

The central 95% of the area under the Z distribution is between -1.96 and +1.96, which can be expressed mathematically as

$$P(-1.96 < Z < 1.96) = 0.95 \qquad \text{(A)}$$

This central 95% region of the Z distribution is shaded below:



If you want to know the confidence interval for the mean of a distribution, $\hat{\mu}$, the first step is to rewrite the *Z* distribution in terms of $\hat{\mu}$. This is accomplished by first subtracting the mean, μ, from $\hat{\mu}$, so that the distribution is centered at 0, just like the *Z* distribution. Then, by dividing by the standard error of the mean, the standard deviation of the distribution becomes 1, just like the *Z* distribution.

$$Z = \frac{\hat{\mu} - \mu}{\sigma / \sqrt{n}} \qquad \text{(B)}$$

Combining equation (A) and (B), along with a little algebra, leads to:

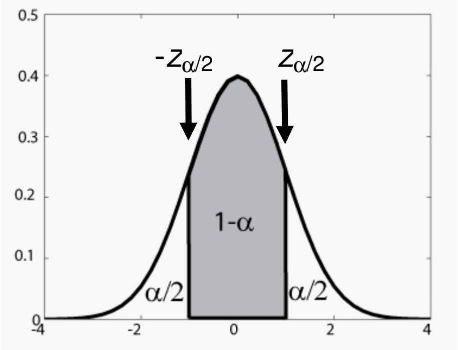
$$P(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

This means that our confidence interval for the mean, μ, is $\hat{\mu} \pm 1.96\frac{\sigma}{\sqrt{n}}$. A confidence interval should be interpreted as follows: the probability that the confidence interval for $\hat{\mu}$ covers the true μ is 95%.

The general form of a 100(1-α)% confidence interval for the mean is written as

$$\hat{\mu} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The number $z_{\alpha/2}$, denotes the boundary of the central 100(1-α)% area of the *Z* distribution, as depicted below. It can be looked up in a table in any stats book.



In order to use the *Z*-distribution, you need to know the true variance of underlying distribution. In reality, this is almost never the case. If you only have an estimate of the variance, you can use the *T*-distribution instead. In that case, the confidence interval for the mean would read as follows:

$$\hat{\mu} \pm t_{\alpha/2,n-1} \frac{\hat{\sigma}}{\sqrt{n}}$$

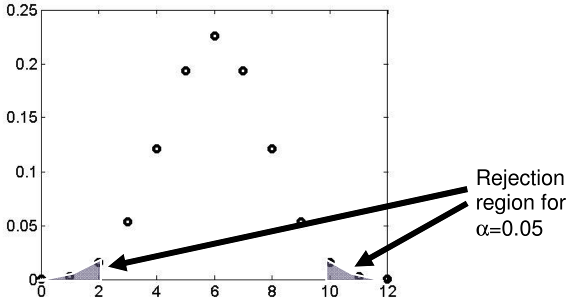
Unlike the *Z*-distribution, the *T*-distribution depends on the sample size *n*. The *T*-distribution looks the same as the *Z*-distribution for large sample size (*n*>25), so in that case it doesn't matter which distribution you use.

Hypothesis Testing

Let's say you are wondering if a coin is fair. You could calculate a point estimate for the probability of heads, and calculate a confidence interval on that probability after flipping the coin, say, 12 times. If the confidence interval does not overlap with 0.5, you may be suspicious that the coin isn't fair. But how do you know if you can reject the hypothesis that the coin is fair? This is where hypothesis testing comes in.

The null hypothesis is that the coin is fair (*p*=.5). The probability distribution that describes the null hypothesis is a binomial distribution with *n*=12 and *p*=.5. The alternative hypothesis is that the coin is not fair (*p*≠.5). How many heads would we need to observe in 12 trials to reject the null hypothesis? The convention is to select a cutoff such that the probability of *incorrectly* rejecting the null hypothesis is less than .05 (α=.05).

In the plot below, we have displayed the null distribution for the probability if seeing a certain # of heads in 12 coin flips. The shaded region on the edges of the distribution is the rejection region for α=.05. The highlighted region is the outer 5% of the curve. If we find between 3 to 9 heads, we cannot reject the null hypothesis. If we find 0-2 or 10-12 heads, we can reject the null hypothesis.



There are 5 key steps to statistical hypothesis testing. These same steps are followed regardless of the specific hypothesis being tested.

1. *Select a statistical test*

   This chart is a list of some common statistical tests:

What you're testing	**Assumptions**	**Test Name**	**Hypotheses**	**Notes**
Something about μ	Normal distribution Variance known	z-test	H_0: $\mu_{test}= \mu_{null}$ H_1: $\mu_{test}\neq \mu_{null}$ or $\mu_{test}> \mu_{null}$ or $\mu_{test}> \mu_{null}$	Use a two-tailed t-test if H_1 is an inequality. Use a one-tailed test if H_1 is directional (i.e. you have previous evidence which lets you rule out one direction).
Something about μ	Normal distribution Variance unknown	t-test	H_0: $\mu_{test}= \mu_{null}$ H_1: $\mu_{test}\neq \mu_{null}$ or $\mu_{test}> \mu_{null}$ or $\mu_{test}> \mu_{null}$	Use a two-tailed t-test if H_1 is an inequality. Use a one-tailed test if H_1 is directional (i.e. you have previous evidence which lets you rule out one direction). # degrees of freedom = n-1
How well does an observed frequency distribution of certain events fit the distribution predicted by the null hypothesis?	*n* is larger than 30. Expected frequencies are all larger than 5. Error is normally distributed.	Pearson's chi-square test	H_0: observed distribution comes from predicted distribution H_1: observed distribution doesn't come from predicted distribution	# degrees of freedom = # categories – 1
Are all of the means from several distributions the same?	Everything is normal and has equal variance	ANOVA	H_0: all μ's are equal H_1: at least one pair of μ's are not equal (but DOESN'T tell you which one)	Based on idea that observed value = average + variation within group + variation across groups. Null hypothesis is that variation across groups is 0 Degrees of freedom within groups = sample size within a group -1 Degrees of freedom between groups = # groups -1

2. *State the hypotheses: null & alternative*

   The null hypothesis is a hypothesis that you are testing. You want to know whether or not you can disprove it, at a particularly confidence level, based on your data. For example, for a t-test, the null hypothesis could be that the mean of your distribution is 0 (μ=0), while the alternative could be that the mean of your distribution is not 0 ($\mu \neq$0).

3. Identify the test statistic under the null hypothesis

The relevant test statistic depends on what hypothesis you are testing. The chart below shows some test statistics for some common statistical tests.

One sample z-test (testing if the mean is different from a value; known variance)	$z = \dfrac{\hat{\mu} - \mu}{\sigma / \sqrt{n}}$
One sample t-test (testing if the mean is different from a value; unknown variance)	$t = \dfrac{\hat{\mu} - \mu}{\hat{\sigma} / \sqrt{n}}$
Two sample t-test (testing for a difference in the mean of 2 populations)	$t = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2_{x1} / n_1 + \hat{\sigma}^2_{x2} / n_2}}$
One proportion z-test (testing if a population proportion is different from a value)	$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$

4. Determine the rejection region of the test statistic for the selected significance level α.

Among all the sets of possible values of the test statistic, we choose the values that represent the most extreme evidence *against* the hypothesis, given some significance level. The significant level, α, which is often set to .05, represents the probability of *incorrectly rejecting* the null hypothesis. Whatever significance level you choose, you can look up the value of the test statistic for that significance level in a table. For instance, as depicted by the shaded region below, if you are using a *z*-test to test the null hypothesis that μ=0, the rejection region would be a test statistic >1.96 or <-1.96. (This follows from the fact that $z_{\alpha/2}$=1.96 for α=.05.)

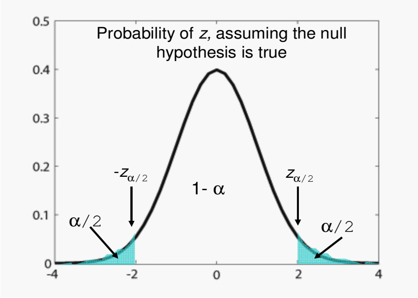
α is the probability of incorrectly rejecting the null hypothesis

Rejection region:

$z > z_{\alpha/2}$

$z < -z_{\alpha/2}$

If α=.05, $z_{\alpha/2}$=1.96



5. Calculate the value of the test statistic for the data set.

Calculate the value of the test statistics based on your data. The table above has the formula for some common test statistics.

6. Determine whether or not the test statistic falls within the rejection region

Compare the value of the test statistic to the rejection region. If it's in the rejection region, you reject the null hypothesis. Otherwise, you can't reject the null hypothesis.

What about if you have **multiple hypotheses** about your data? For example, you want to know about the following things:

1) did the dendrite change in size?
2) did the axon change in size?
3) are there more presynaptic terminals?
4) are there more postsynaptic terminals?

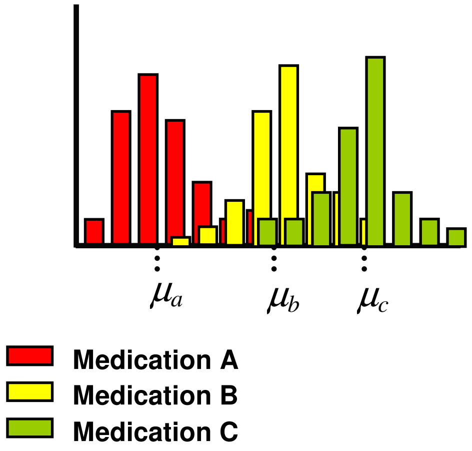
One possibility would be to test each hypothesis, one at a time. But this would cause an error. If you set α=.05 for each hypothesis, and you have 4 independent tests, you have a greater than 5% chance of incorrectly rejecting at least one of the 4 hypotheses. If they're all actually true, you actually have a 18% chance to incorrectly reject one of the null hypotheses!

An extremely simple correction to overcome this multiple hypothesis testing issue is the "Bonferroni correction". You divide the desired α level **for the whole test** by the number of tests to get the corrected level of α **for each test**.

$\alpha_{\text{bonf}} = \alpha/n$

There are many more sophisticated corrections that can also be used to take into account multiple hypotheses. These tests exist in most statistical analysis packages, and are preferable to the Bonferroni correction because they are less conservative. These less conservative adjustments will result in a larger adjusted significance levels α, and therefore you'll be more likely to end up with significant results.

A common statistical test that can be useful if you have multiple situations that you are comparing is the ANOVA (analysis of variance). It tests for effect of 1 or more categories (or “factors”) on the mean value of the data. Each category may include 2 or more conditions (or “groups”).



1-factor ANOVA tests whether *all* group means are the same.
Null hypothesis: $\mu_a = \mu_b = \mu_c$
Alternative hypothesis: At least one pair is different.

In ANOVA, the total variance in the data is partitioned into two components:
across-groups: variance of the data across groups
within-groups: variance of the data within groups

If the across group variance is sufficiently larger than the within group variance, you reject the null hypothesis that all means are equal. The ANOVA relies on the *F*-test, which tests for differences in the ratio of variances.

For 2 factor ANOVA, there are 2 different factors that you are interested in. For example: How does rodent lifespan depend on exercise and medication?

		Exercise	
		0 hr	1hr
Medication	A	2.5	3.0
	B	2.7	2.7
	C	5.0	6.3

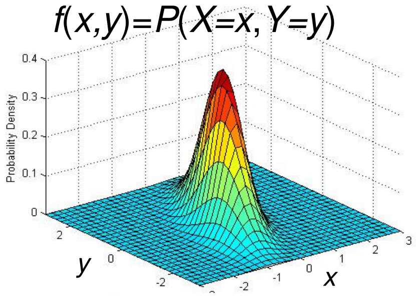
The things the 2-way ANOVA would be testing are:

- Does exercise effect lifespan?
- Does medication effect lifespan?
- Is there an interaction between medication and lifespan? (*Interaction* means that you cannot simply sum the influence of exercise and lifespan to predict the lifespan. In other words, medication affects lifespan differently, depending on the exercise level.)

In summary, we use an ANOVA to decide whether we can reject the null hypothesis that all group means are equal. An important thing to keep in mind is that even if we reject the null hypothesis, we still don't know which pairs of means are different from one another. Usually, to figure out which means are different, people perform multiple hypothesis testing on the factor that ANOVA identifies as "significant".

Independence & covariance

So far, we have discussed the probability distribution of a SINGLE random variable. You can also have a probability distribution for multiple random variables, which is known as a *joint probability distribution*. The sample probability distribution below depends on both *X* and *Y*. In this case, you are most likely to observe a value of both *X* and *Y* near 0 (because the distribution peaks at 0 for both variables).



What's the probability distribution of *X*, given you already know the value of *Y?* This probability is written mathematically as *P*(*X*=*x*|*Y*=*y*), and is known as a *conditional probability*. For instance, let's say you have a deck of cards. The probability you draw a spade is ¼, i.e. *P*(draw a spade)= ¼ . But what's the conditional probability you draw a spade, given that you know the card is black?

P(draw a spade|black card) = ½

In order to calculate a conditional probability, you can use the following equation:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

In words: the probability of *X* given *Y* is the probability of *X* and *Y*, divided by the probability of *Y*.

As an example, consider the following joint probability distribution for *X* and *Y* *P*(*X*=*x*,*Y*=*y*):

		Y			
		0	**1**	**2**	**3**
X	**0**	1/8	2/8	1/8	0
	1	0	1/8	2/8	1/8

Let's calculate $P(X=1|Y=1)$. Based on the chart, $P(X=1,Y=1)$=1/8 and P(Y=1)=2/8+1/8=3/8. Plugging in for Bayes' rule, we end up with $P(X=1|Y=1)$=1/3. In other words, if someone told us that *Y* is 1, then *X* has a 1/3 chance of being 1.

Two random variables *X* and *Y* are statistically *independent* if knowledge of the value of *X* does not affect the probability distribution of *Y* (and vice versa)

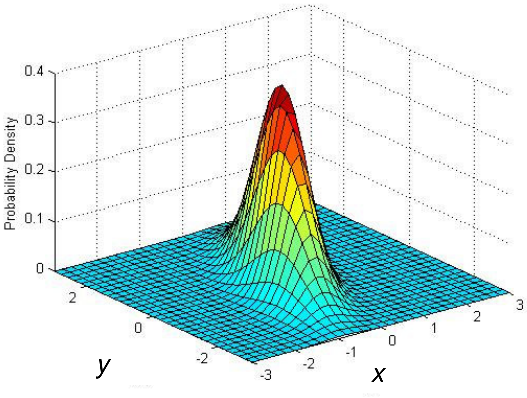
Mathematically, *X* and *Y* are independent if the following statements are true.

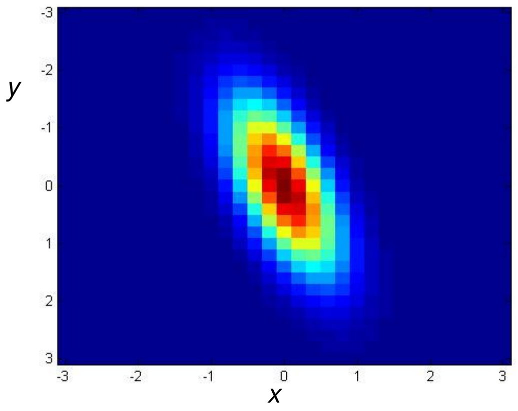
$P(X=x|Y=y) = P(X=x)$

or

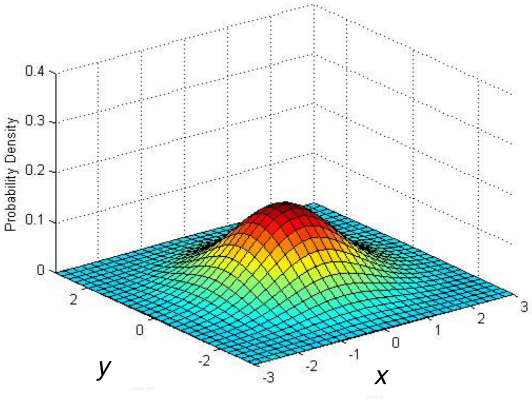
$P(X=x,Y=y) = P(X=x)\ P(Y=y)$

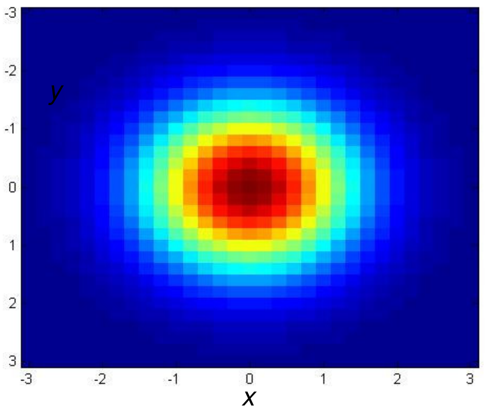
In the figure below, *X* and *Y* are ***dependent***. That's because if you know *Y*, that will change the probability distribution for *X*.





On the other hand, in the figure below, *X* and *Y* are ***independent***. That's because if you know *Y*, that does *not* change the probability distribution for *X*.

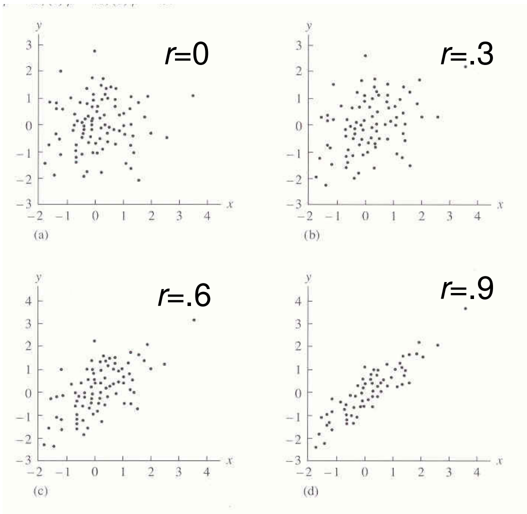




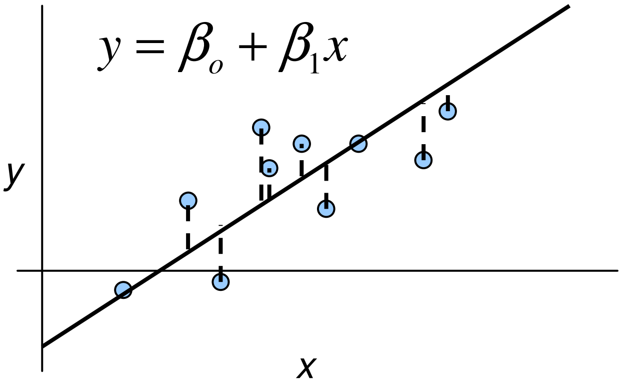
Correlation coefficient and regression

Another way of describing the relationship between two variables is the ***covariance***. It's a measure of how much two variables change together. If *X* is larger than average when *Y* is larger than average, and *X* is smaller than average when *Y* is smaller than average, the covariance is *positive*. On the other hand, if *X* is smaller than average when *Y* is larger than average, and *X* is larger than average when *Y* is smaller than average, the covariance is *negative*. If they are independent, and *X* and *Y* have no relation to each other, and the covariance is 0.

The ***correlation coefficient*** is closely related to the covariance; it's just the covariance normalized by the variance (so that it ranges from -1 to 1). It indicates the strength and direction of a linear relationship between *X* and *Y*. In the plot below, the variable *r* refers to the correlation coefficient.



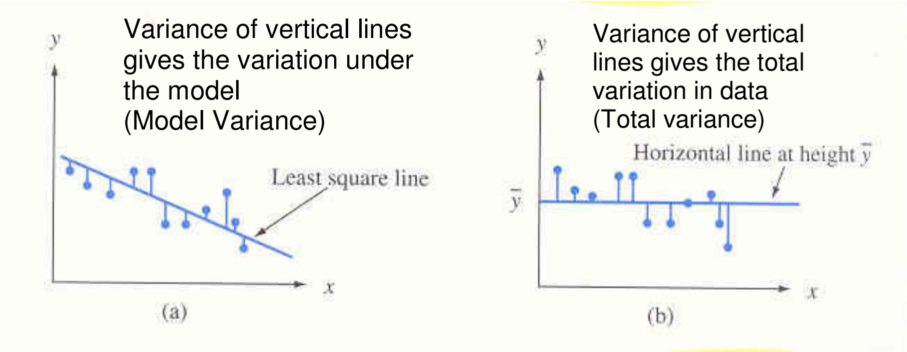
The ***linear regression*** is the line that minimizes the "least square error": the sum of squared deviations between the data and the fitted line.



How does the correlation coefficient relate to a linear regression? If r is the correlation coefficient, then r^2 is the proportion of observed y variation that can be explained by a linear regression model. Mathematically,

r^2 = 1 – (Model Variance)/(Total Variance)

(See plot below for definitions of Total and Model variance.) Often, people report the value of r^2 after fitting a regression model. If r^2 is near 1, the model is a good fit (it explains most of the variance in the data), while if r^2 is near 1, you need to find a new model, because it is not capturing much of the variance in the data.



When is it appropriate to calculate the correlation coefficient, and when do you take a linear regression?

- Use correlation coefficient to determine the strength of a linear association between *X* and *Y*.
- Use linear regression when you want to use *X* as a predictor for *Y*. Unlike correlation, it's NOT symmetric between *X* and *Y*.