MATH CLUB 2/28/2013

Note Title 2/28/2013

THE BEAUTY OF STATISTICAL TESTS

(1) Normal distribution

Random variable X prob.

Example: flip a coin $X = 1$ if heads $1/2$

$X = 0$ if tails $1/2$

If X takes on any real value:

usually, probability $X =$ given number

$$P\{X = x\} \overset{\text{is } 0}{=} 0$$

Instead, we have a density function

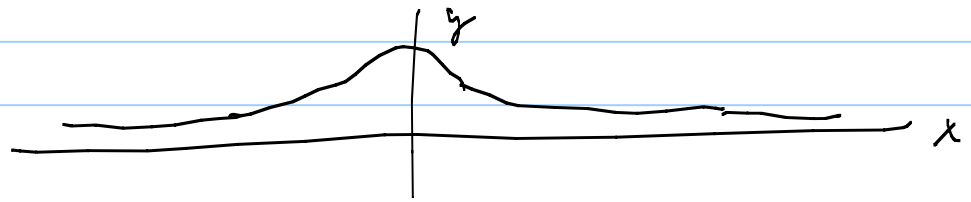
= a ≥ 0 function on $\mathbb{R}$, say, $f(x)$.

$$P\{a \leq X \leq b\} = \int_a^b f(x)\,dx$$

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

density function $\simeq$ distribution

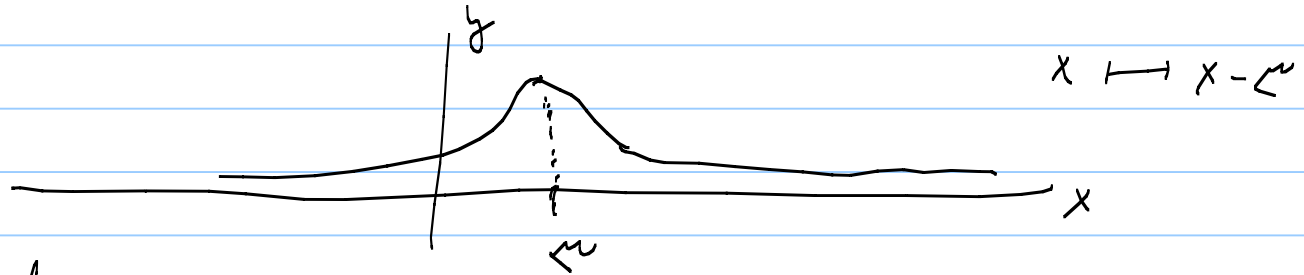
The normal distribution



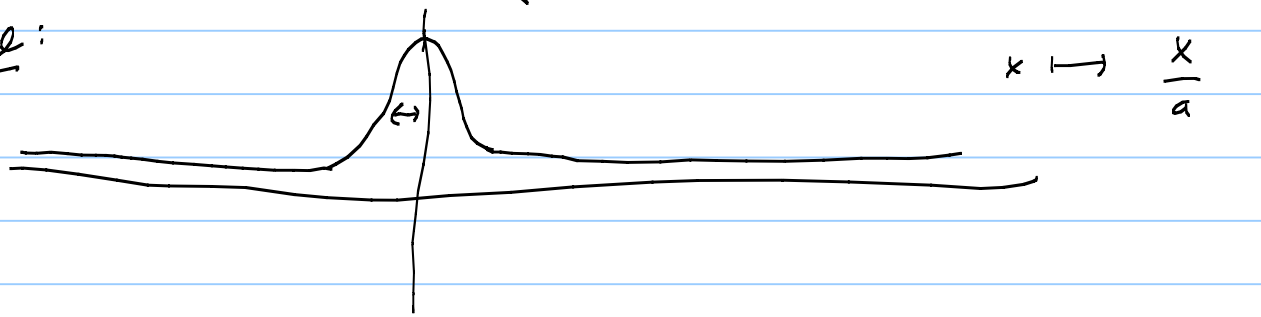
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Why is the normal distribution important?

"Obvious transformations": shift



scale:



The more general normal distribution:

$$\frac{1}{a\sqrt{2\pi}}\, e^{\frac{(x-\mu)^2}{2a^2}}$$

μ = average
= expected value
= $E(X)$

a is a measure of the width of the curve

a^2 = variance $\quad var(X) = E\left((X - E(X))^2\right)$.

The importance of normal distribution:
it does not change (up to changing these two parameters μ, a) when I add independent variables.

Suppose I have two random variables X, Y.

I can talk about a joint density $f(x,y)$

$X = x$ & $Y = y$

$$P\{(X,Y) \in S\} = \int_S f(x,y)\,dx\,dy.$$

Independent means $f(x,y) = g(x)\,h(y)$

individual densities

Key fact: A sum of two independent normal variables is normal.

Application: A sum of a large number of independent random variables with the same distribution, it is approximately normal.

↑ scale to get finite average and variance.

Informal argument: If such a limit exists, the mystery distribution must satisfy the "key fact":

$$\underbrace{X_1 + \cdots + X_m}_{\sim \text{ mystery distribution}} + \underbrace{X_1 + \cdots + X_n}_{\sim \text{ mystery distribution}}$$

variance prop. to m — variance prop. to n

This uniquely characterizes the normal distribution!

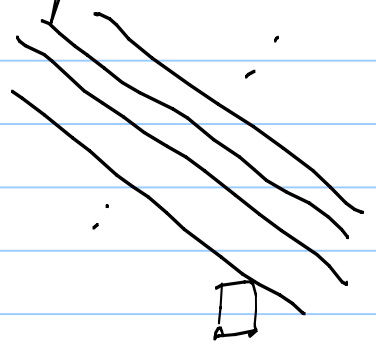
Proof sketch of key fact: (think averages = 0)

$$\cdots \int e^{-\frac{x^2}{2a^2}} \cdot e^{-\frac{(z-x)^2}{2b^2}} \, dx$$

examine dependence on z.

$$e^{-(\alpha x + \beta z)^2 - \gamma z^2}$$

↑ bringing to perfect square.



---

(2) Multivariable normal distribution

density

e^{-} quadratic expression

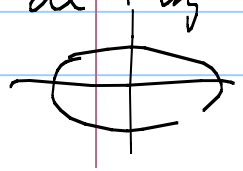
negative definite

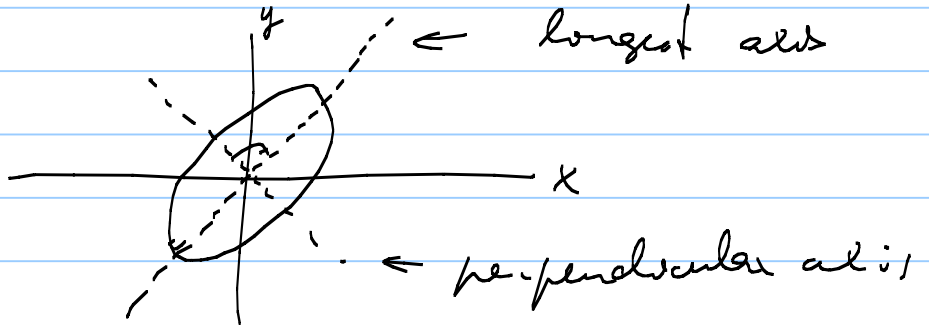
$\lim_{\to\infty} = -\infty$

two variable: what are the curves on which it is constant?

Ellipses:

$ax^2 + by^2$



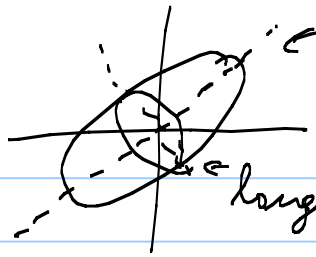


← longest axis

← perpendicular axis

>2 variables

Ellipsoid



← longest axis

← longest perpendicular axis

etc.

⇒ After turning to perpendicular axes,

The normal distribution

= (, ⋯ ,) independent normal distributions

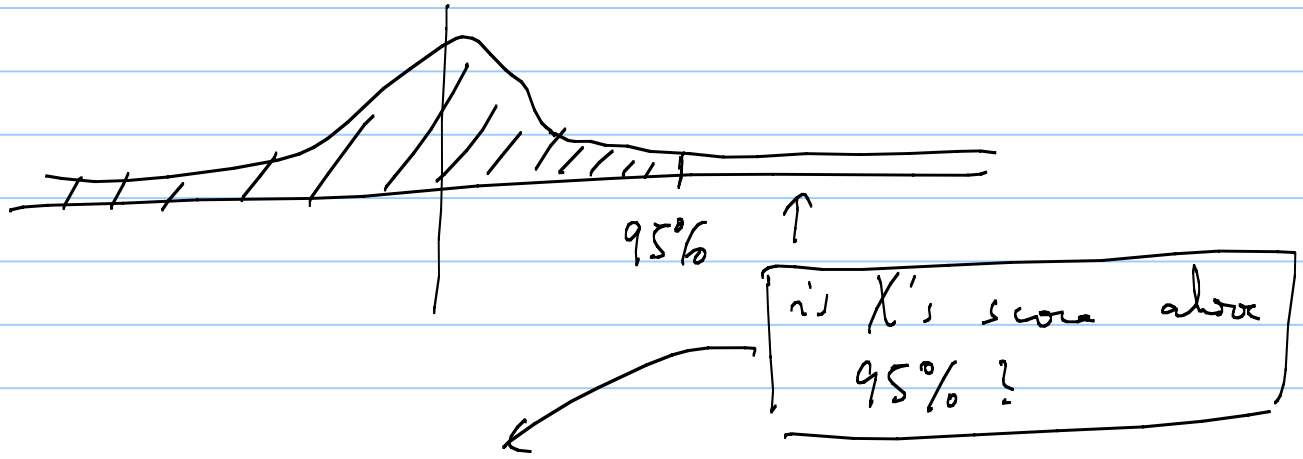
---

(3) What is a statistical test?

There is a tutor J. Is this tutor effective?

Take a random student X. Give X a lesson with tutor J. Let X take the test.

Score distribution on the test:



Recipe: If YES → say: "Yes, the factor J is effective"

If NO → say nothing.

I/J effective? \ I: X > 95%	Yes	No
Yes	Right	Right
→ No null hypothesis	Wrong ≤ 5%	Right

J not effective: 5%

Bayes analysis

You can only disprove, but never have a null hypothesis.

④ The Pearson χ^2 test

Draw N balls with return. k colors, null hypothesis

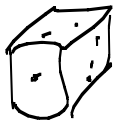
color i occurs with probability p_i.

We get a_i balls of color i.

$$\chi^2 = \frac{(a_1 - p_1 N)^2}{p_1 N} + \frac{(a_2 - p_2 N)^2}{p_2 N} + \cdots \frac{(a_k - p_k N)^2}{p_k N}.$$

$\chi^2 \sim \sum_{k-1}$ independent standard normal variables

Example: Cast a die 30 times.



Get		
	10 times	1
	10 times	2
	5 times	3
	1 times	4
	never	5, 6.

Is the die biased?

$p_i \cdot N = 5$

$$\chi^2 = \frac{(10-5)^2}{5} + \frac{(10-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(0-5)^2}{5} + \frac{(0-5)^2}{5}$$

$$= 20$$

χ^2 with $k-1=5$ degrees of freedom

95% 11.07

$14.4 > 11.07$ Yes!

The die is 95% certainly biased

Sketch proof that χ^2 is distributed as I said.

(1) ~ multivariable normal distribution.

$N >> 0$

Which quadratic expression do I have?

covariance

$$\operatorname{cov}(X,Y) = E\big((X - E(X))(Y - E(Y))\big)$$

Make covariance matrix; look for principal axes:

	a_1	----	a_k
a_1	$1-p_1$	$-\sqrt{p_1 p_2}$	$---$ $-\sqrt{p_1 p_k}$
⋮	$-\sqrt{p_1 p_2}$	$1-p_2$	$-\sqrt{p_i p_j}$
a_k			$1-p_k$

$=$

$$= \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_k} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{p_1} & \cdots & \sqrt{p_k} \end{pmatrix}$$

↑ vector of length 1

if this vector were

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

then $\simeq \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ & & & 0 \end{pmatrix}$

$k-1$ independent standard normal variables! $\square$