Hypothesis testing

We have talked about point estimation, say estimating μ.

and bit about confidence intervals (like $P(\bar{x}-\frac{1.96\sigma}{\sqrt{N}} \leq \mu \leq \bar{x}+\frac{1.96\sigma}{\sqrt{N}}) = 95\%$), which is related to interval estimation.

In hypothesis testing, we are saying yes or no to some hypothesis, when presented with an alternative hypothesis.

$x \sim$ normal variable, σ known

$H_0: \mu = \mu_0$ $\qquad$ $H_1: \mu > \mu_0$

(In fact this is a whole class of hypothesis)

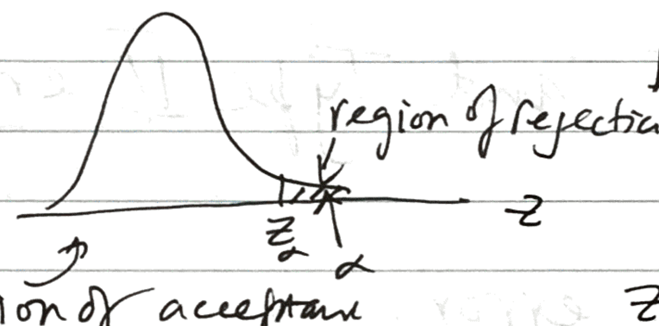
Traditionally we define a region of acceptance for a test statistic. If the statistic observed falls in the region H_0 is accepted, Otherwise, it is rejected

For example, if we knew σ (a tall order),

$\bar{X} = \frac{X_1 + \ldots + X_N}{N}$ is distributed as $N(\mu_0, \frac{\sigma^2}{N})$

according to H_0. Let $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$. $Z \sim N(0,1)$ then

$P(Z \geq z_\alpha \mid H_0) = \alpha$



Region of acceptance

$Z \in (-\infty, z_\alpha]$

Region of rejection complimentary to that.

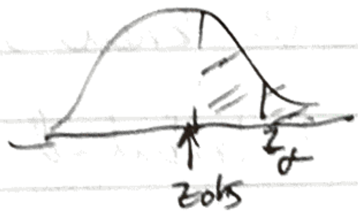
This is a one tailed one sample z-test.

The prob. α is called level of significance.

In this version you get a yes/no answer.

Alternative: Convert z_{obs} to a p-value

$P(Z \geq z_{obs} \mid H_0)$. If this value is greater than α, then accept.



Type I and Type II errors.

Type I error.

H_0 true but rejected

Type II error

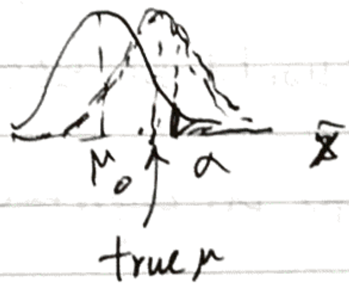
H_a true but H_0 accepted.

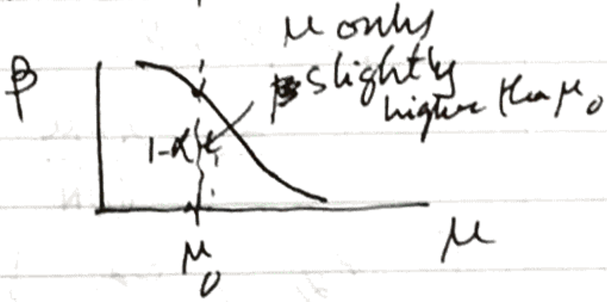
Prob of Type I error $\rightarrow \alpha$

Prob of type II error $\rightarrow \beta$

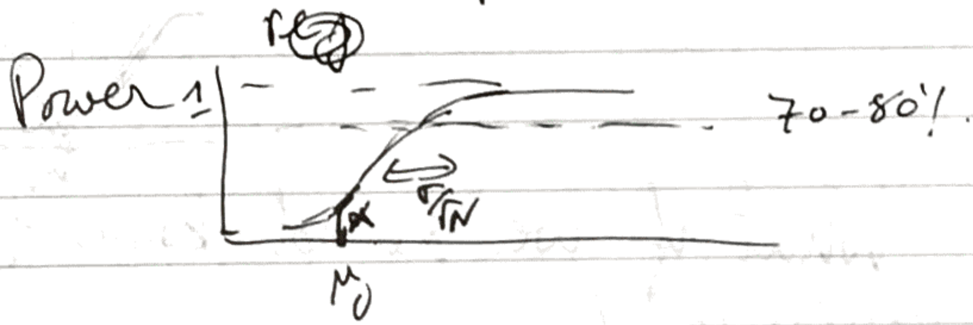
is more complicated.

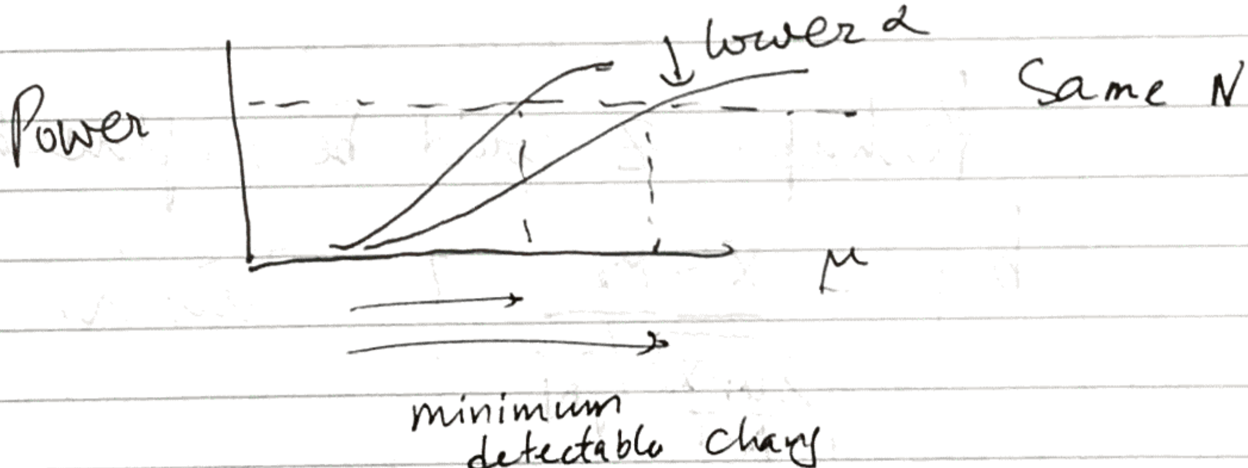
β depends on true μ

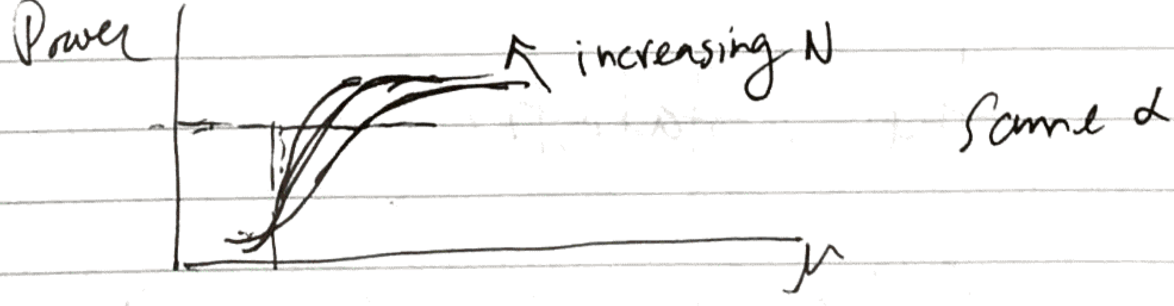




Power $= 1-\beta =$ Prob if Ha is true Ho would be rejected



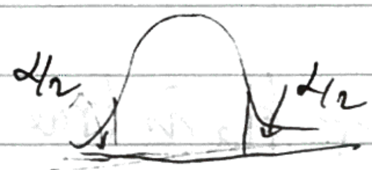




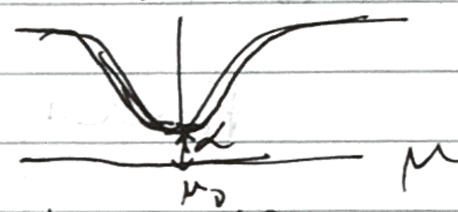
Two sided tests for $\mu = \mu_0$
vs $\mu \neq \mu_0$

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$ should be ~~between~~ in the interval

$[-z_{\alpha/2}, z_{\alpha/2}]$ for acceptance.



Power

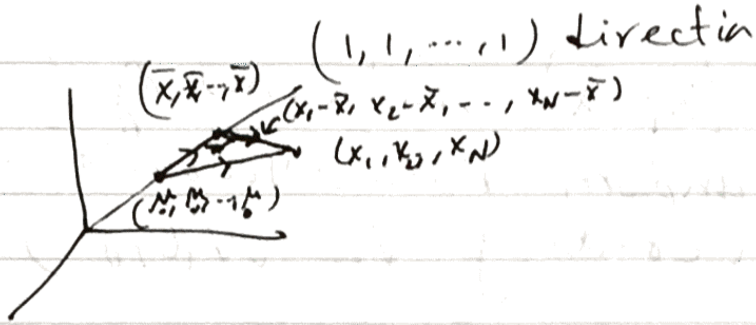


What if we do not know the variance?

Perhaps z could be replaced by $\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}/\sqrt{N}}$? This is

the t statistic.

Quite remarkably, we know the exact distribution of t.



Consider the vector $(X_1 - \mu_0, \dots, X_N - \mu_0)$
with entries i.i.d $\mathcal{N}(0, \sigma^2)$

It's projection along the $(1, 1, 1, \dots)$ direction
is obtained by taking the dot product
with $\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}}\right)$.

That dot product is $\frac{X_1 + \dots + X_N - N\mu_0}{\sqrt{N}}$

$$= \frac{N(\bar{X} - \mu_0)}{\sqrt{N}} = \sqrt{N}(\bar{X} - \mu_0)$$

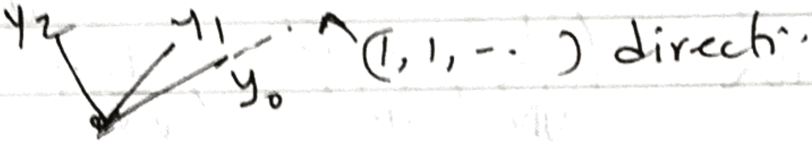
The component perpendicular to $(1, 1, \dots)$
lives in an $N-1$ dimensional space
call $N-1 = \nu$. That is the degree
of freedom of $X_i - \bar{X}$ components.

With a change of coordinates, we can write

$(x_i - \bar{x})$ in terms of $\nu(=N-1)$ independent random normal variables distributed according to $\mathcal{N}(0,\sigma^2)$.

$$\prod_i e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \sim e^{-\frac{y_0^2}{2\sigma^2}} \, e^{-\sum_{j=1}^{\nu} y_j^2/2\sigma^2}$$

$$Y_0 = \sqrt{N}\,(\bar{x} - \mu_0)$$



$$\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{j=1}^{\nu} y_j^2$$

So $$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{N-1}}/\sqrt{N}} = \frac{Y_0}{\sqrt{\sum y_j^2/\nu}}$$

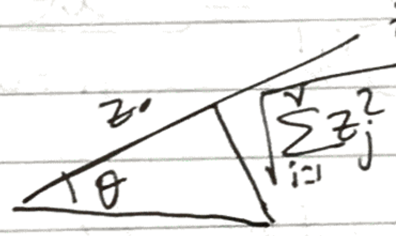
We could write $Y_0 = \sigma z_0$, $y_j = \sigma z_j$

and z_0, z_j would all be independent standard normal variates.

So $t = \dfrac{z_0}{\sqrt{\sum_{j=1}^{\nu} z_j^2 / \nu}}$

$\sum_{j=1}^{\nu} z_j^2$ is called a χ^2 variate of degree of freedom ν

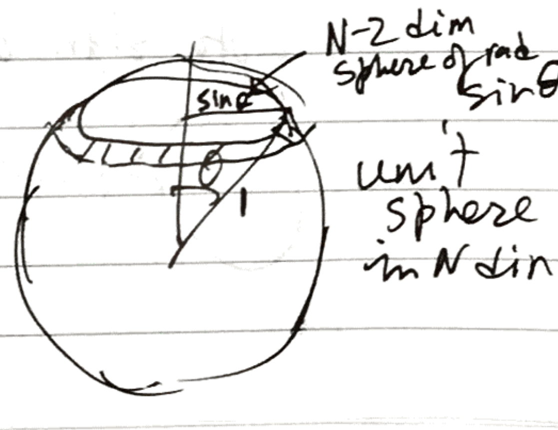
[Sum of squares of ν independen ~~norm~~ standard normal variates.]



$$\frac{z_0}{\sqrt{\sum_j z_j^2}} = \cot\theta = \frac{t}{\sqrt{\nu}}$$

It depends only on a random angle

In $N = \nu + 1$ dimension θ distribution goes as $\sin^{\nu-1}\theta \, d\theta$



Vol $\sim \sin^{N-2}\theta \, d\theta$

$\sim \sin^{\nu-1}\theta \, d\theta$

$$1+\cot^2\theta = \frac{\sin^2\theta+\cos^2\theta}{\sin^2\theta} = \frac{1}{\sin^2\theta}$$

So $\sin\theta = \dfrac{1}{\sqrt{1+\cot^2\theta}}$

$$d\cot\theta = -\operatorname{cosec}^2\theta\, d\theta$$
$$= -\frac{d\theta}{\sin^2\theta} = -(1+\cot^2\theta)\, d\theta$$

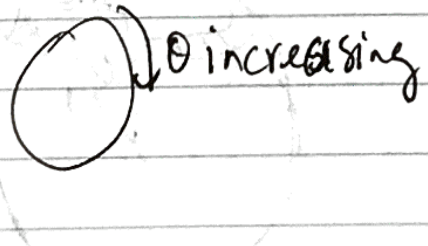
$$\sin^{\nu-1}\theta\, d\theta = (1+\cot^2\theta)^{\frac{1-\nu}{2}} d\theta$$

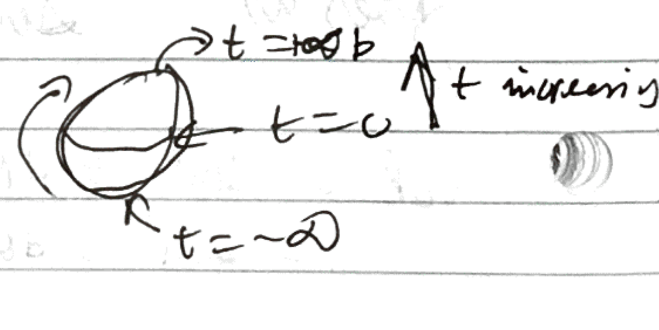
$$= +\frac{(1+\cot^2\theta)\, d\theta}{(1+\cot^2\theta)^{\frac{\nu+1}{2}}} = -\frac{d(\cot\theta)}{(1+\cot^2\theta)^{\frac{\nu+1}{2}}}$$

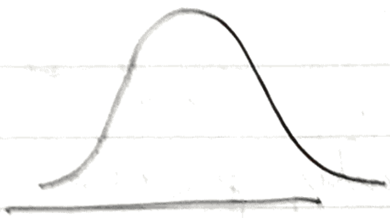
Changing to $\cot\theta = t/\sqrt{\nu}$

$$\sin^{\nu-1}\theta\, d\theta \longrightarrow \frac{dt}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

upto sign and constants.







$$C_1 \int_0^{\pi} d\theta \sin^{\nu-1}\theta = C_2 \int_{-\infty}^{\infty} \frac{dt}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

← Powerlaw tail $\frac{z_0}{\sqrt{z_i^2}}$ ← The χ^2 variable fluctuate and can be small occasionally.

We can do one and two sided test with t.

If we have two samples of size n_1, n_2 and we know σ_1, σ_2, a two sample z test could deside if the means are the same or not. $H_0: \mu_1 = \mu_2$, $H_a: \mu_1 > \mu_2$ or $\mu_1 \neq \mu_2$

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}}$$

The corresponding thing for t test is tricky.

Paired t-test.

If ~~the~~ $n_1 = n_2$ and the x's are naturally paired, take difference

$$d_i = x_{1i} - x_{2i}$$

Do a t-test on d_i mean being zero or not.

Pooled t-test

If we don't know σ's but $\sigma_1 = \sigma_2$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is a t statistic

of degree $n_1 + n_2 - 2$,

where $S_p = \sqrt{\frac{\sum_i (X_{1i} - \overline{X}_1)^2 + \sum_j (X_{2j} - \overline{X}_2)^2}{n_1 + n_2 - 2}}$

Two sample-t test

This is an approximate test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$S_\alpha^2 = \sum_{i=1}^{n_\alpha} (X_{\alpha i} - \bar{X}_\alpha)^2 / (n_\alpha - 1) = \frac{\sigma_\alpha^2}{n_\alpha - 1} \sum_{j=1}^{n_\alpha - 1} Z_{\alpha j}^2$$

We need to know how does $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ fluctuate compared to the mean.

For $\sum_{j=1}^{\nu} Z_j^2$ $\qquad E\left[\sum_{j=1}^{\nu} Z_j^2\right] = \nu$

$$\mathrm{var}\left[\sum_{j=1}^{\nu} Z_j^2\right] = \sum_{j=1}^{\nu} \mathrm{var}[Z_j^2] = 2\nu$$

$$\mathrm{var}[Z^2] = E[Z^4] - E[Z^2]^2 = 3 - 1^2 = 2$$

$$\frac{2E[\chi_\nu^2]^2}{Var[\chi_\nu^2]} = \nu$$

If we compute that for

$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$, we get

Expectation $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Var $\frac{2\sigma_1^4}{(n_1-1)n_1^2} + \frac{2\sigma_2^4}{(n_2-1)n_2^2}$

So we should use $\frac{2\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{2\sigma_1^4}{(n_1-1)n_1^2} + \frac{2\sigma_2^2}{(n_2-1)n_2^2}}$

as the effective d.f.

We use its estimate

$$d.f. \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^4}{(n_1-1)n_1^2} + \frac{S_2^4}{(n_2-1)n_2^2}\right)}$$

Enough about means, what about hypotheses regarding variances?

Say X_i iid $N(\mu, \sigma^2)$ $\quad i=1,\dots,N$

Remember that $\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{j=1}^{\nu} Y_j^2$ with $\nu = N-1$. If we set $Y_j = \sigma Z_j$ then Z_j's are i.i.d $N(0,1)$.

The hypothesis $\sigma = \sigma_0$ against some alternative hypothesis could be decided by the statistic

$$\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{\sigma_0^2}$$

To test these hypotheses we need to the distribution of $\chi^2 = \sum_{j=1}^{\nu} Z_j^2$

$$P(z)\, d^\nu z \sim e^{-\sum_j z_j^2/2}\, dz_1 \cdots dz_\nu$$

Let $\sum_{j=1}^{\nu} z_j^2 = \chi^2$. The radius in the z space is χ.

$P(z_1, \ldots, z_\nu) \to e^{-\chi^2/2}$

If I go to polar coordinates

$$dz_1 \cdots dz_\nu \to \chi^{\nu-1}\, d\chi\, d\Omega^{\nu-1}$$

angular degrees of freedo

Integrate out Ω's

$$C\, e^{-\chi^2/2} \chi^{\nu-1} d\chi$$

Normalizign $\displaystyle\int_0^\infty e^{-\chi^2/2} \chi^{\nu-1} d\chi$

$$= \int_0^\infty e^{-y} (2y)^{\frac{\nu-2}{2}} dy = 2^{\frac{\nu-2}{2}} \int_0^\infty y^{\frac{\nu}{2}-1} e^{-y} dy$$

$$= 2^{\frac{\nu-2}{2}} \Gamma\left(\frac{\nu}{2}\right)$$

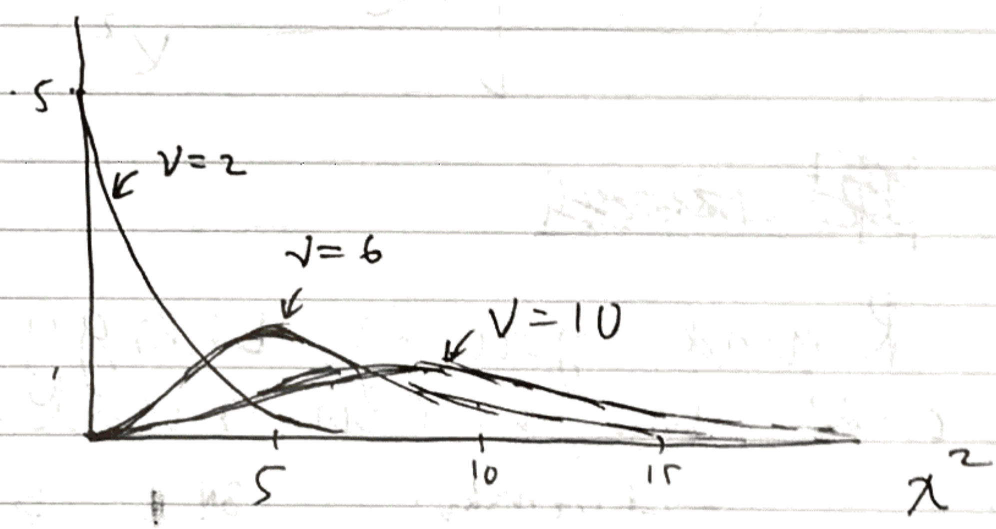
So $C = \dfrac{1}{2^{\frac{\nu-2}{2}} \Gamma\left(\frac{\nu}{2}\right)}$

If we call $\chi^2 = x$

$$\frac{1}{2^{\frac{\nu}{2}-1}\,\Gamma\left(\frac{\nu}{2}\right)}\, e^{-\chi^2/2}\, \chi^{\nu-1}\, d\chi$$

$$= \frac{1}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)}\, e^{-\chi^2/2}\, \left(\chi^2\right)^{\frac{\nu-2}{2}}\, d\chi^2$$

$$\underbrace{= \frac{1}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)}\, e^{-\frac{x}{2}}\, x^{\frac{\nu}{2}-1}\, dx}_{x(=\chi^2)\ \text{pdf.}}$$



Mode for X: take log and diff $\frac{d}{dx}\left(-\frac{x}{2}+\left(\frac{\nu}{2}-1\right)\log x\right)=0$

$\Rightarrow x_m = \nu - 2$

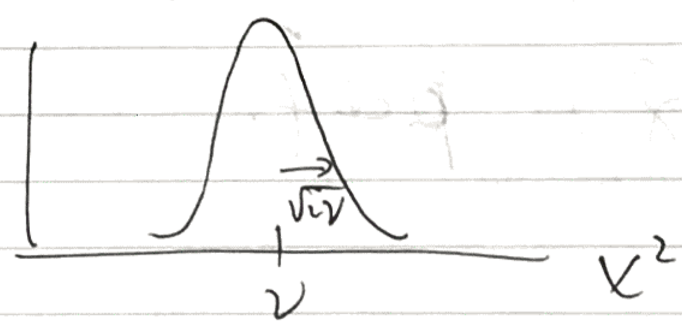
Since $\chi^2 = \sum_{i=1}^{\nu} z_i^2$

$E[\chi^2] = \nu$

$Var[\chi^2] = 2\nu$

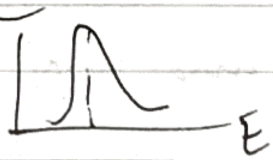
So for large ν, we expect

$\frac{\chi^2 - \nu}{\sqrt{2\nu}}$ to behave like a standard normal variate



Reminds you of energy of ideal gas in canonical ensemble

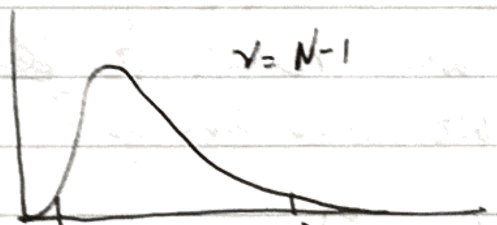
$$e^{-\sum_i p_i^2/2mk_BT} \rightarrow E^{\frac{3N}{2}-1} e^{-\beta E}\, dE$$



$\bar{E} \simeq \frac{3N}{2\beta} = \frac{3Nk_BT}{2}$ = Equipartition law.

Back to practical things!

First we can generate confidence intervals of σ.



$$\mathrm{Prob}\left[\chi^2_{0.025} < \frac{\sum(x_i-\bar{x})^2}{\sigma^2} < \chi^2_{0.975}\right] = 95\%.$$

Let us call $s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$

$$\mathrm{Prob}\left[\frac{(N-1)s^2}{\chi_{0.975}} < \sigma^2 < \frac{(N-1)s^2}{\chi^2_{0.025}}\right] = 95\%.$$

Similarly we could do hypothesis testing.

$H_0 : \sigma = \sigma_0 \quad H_1 : \sigma > \sigma_0$

Reject H_0 if
$$\frac{(N-1)s^2}{\sigma_0^2} > \chi^2_\alpha .$$

What if we need to compare two variances?

i.i.d $X_{1i} \sim N(\mu_1, \sigma_1^2)$, $i = 1, \dots, N_1$,
i.i.d $X_{2j} \sim N(\mu_2, \sigma_2^2)$, $j = 1, \dots, N_2$,

$H_0: \sigma_1 = \sigma_2$ $\quad\quad$ $H_a: \sigma_1 > \sigma_2$

Note that
$$\frac{\sum_{i=1}^{N_1}(X_{1i}-\bar{X}_1)^2/(N_1-1)}{\sum_{j=1}^{N_2}(X_{2j}-\bar{X}_2)^2/(N_2-1)} = \frac{S_1^2}{S_2^2} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = F$$

Where χ_1^2, χ_2^2 are independently distributed χ^2's with d.f. $\nu_1 = N_1 - 1$, $\nu_2 = N_2 - 1$.

What is the distribution of F for degrees of freedom ν_1, ν_2

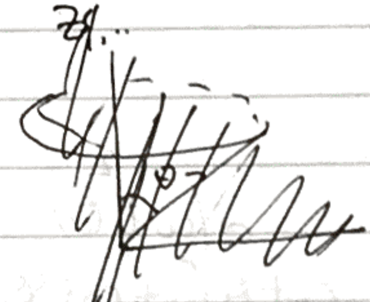
Construct standard normal variables

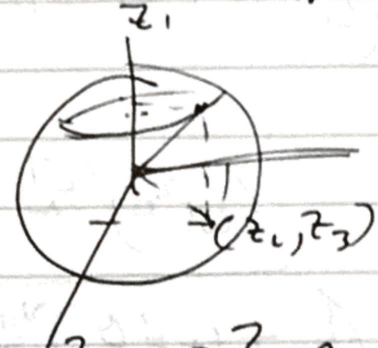
$Z_1, \ldots, Z_{\nu_1}, Z_{\nu_1+1}, \ldots, Z_{\nu_1+\nu_2}$

$Z_1^2 + \cdots + Z_{\nu_1}^2 = \chi_1^2$

$Z_{\nu_1+1}^2 + \cdot + Z_{\nu_1+\nu_2}^2 = \chi_2^2$ $\qquad F \propto \chi_1^2 / \chi_2^2$

has to do with angle





Consider $\chi_1^2 + \chi_2^2 = \chi^2$. χ is the $\nu_1+\nu_2$ dim radius

$\chi_1 = \chi \cos\theta$

$\chi_2 = \chi \sin\theta$

Phase space $\sim \chi_1^{\nu_1-1} d\chi_1 \, \chi_2^{\nu_2-1} d\chi_2$

$\sim \chi_1^{\nu_1-1} \chi_2^{\nu_2-1} \, d\theta \, d\chi$

χ magnitude does not matter

$\cos^{\nu_1-1}\theta \, \sin^{\nu_2-1}\theta \; d\theta$ [Note that the t distr$^{\underline{n}}$ calculation had $\nu_1 = 1$, $\nu_2 = N-1$]

Also $\frac{4F}{\nu_2} = \left(\frac{\cos\theta}{\sin\theta}\right)^2 = \cot^2\theta$

Once more $\cot^2\theta + 1 = \frac{1}{\sin^2\theta}$

$$\left(\frac{\cot^2\theta}{\cot^2\theta+1}\right)^{\frac{\nu_1}{2}-1}\left(\frac{1}{\cot^2\theta+1}\right)^{\frac{\nu_2}{2}-1} d\theta$$

$$d(\cot\theta) = -(1+\cot^2\theta)\, d\theta$$

So $$\frac{(\cot^2\theta)^{\frac{\nu_1}{2}-1}}{(1+\cot^2\theta)^{\frac{\nu_1+\nu_2}{2}-1}} \frac{d(\cot^2\theta)}{\cot\theta(1+\cot^2\theta)}$$

[redacted] $$\propto \frac{F^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1 F}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}\, dF$$

$$\propto \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_1 F+\nu_2)^{\frac{\nu_1+\nu_2}{2}}}\, dF$$

I will leave findind the normalization constant using [redacted] Beta [redacted] function as an [redacted] exercise.

Note that the mode is at

$$\left(\frac{\nu_1}{2}-1\right)\frac{1}{F_m} = \frac{\nu_1+\nu_2}{2}\,\frac{\nu_1}{\nu_1 F+\nu_2}$$ [redacted]

$\Rightarrow \quad \frac{\nu_1 F_m}{\nu_1 F_m+\nu_2} =$ [redacted] $\frac{\nu_1-2}{\nu_1+\nu_2}$

$\Rightarrow \quad \frac{\nu_1 F_m}{\nu_2}$ [redacted] $= \frac{\nu_1-2}{\nu_2+2} \quad \Rightarrow \quad F_m = \frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+1)}$

For large ν_1, ν_2 F_m [redacted] $\rightarrow 1$

we expect that since $\frac{\chi_1^2}{\nu_1}$ [redacted] and $\frac{\chi_2^2}{\nu_2}$ would both self average to 1; making the ratio F to be 1 [redacted] as well