Back to Normal

Some results in Probability and Statistics

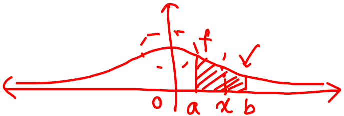
Aniruddha Deb

IIT Delhi

September 24, 2021

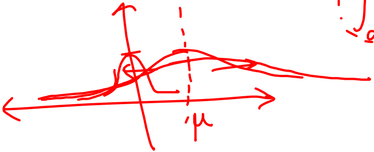
Prerequisites

- Normal Random Variables

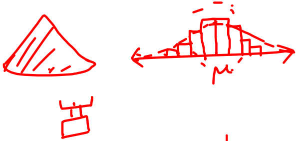


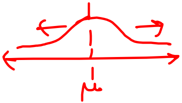
$$f = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\int_a^b f\,dx \Rightarrow \int_{-\infty}^{\infty} f\,dx = 1$$



$$X \sim N(\mu, \sigma^2)$$





Central limit theorem

Prerequisites

- Normal Random Variables
- Moment generating functions

$$M_X(t) = E\left(e^{tX}\right)$$

$$= \int_{-\infty}^{\infty} \underbrace{e^{tx} f(x)\, dx}_{}$$

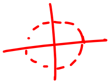
need not converge!!

$\forall t$

$(-h, h)$

Prerequisites

- Normal Random Variables
- Moment generating functions
- Characteristic functions $\rightarrow$ $\phi_X(t) = \left| E\left(e^{itX}\right) \right| \quad \underline{\underline{< 2}}$



$$= \int_{-\infty}^{\infty} e^{itx} f(x)\, dx \rightarrow \forall t$$

* Two RV's are identically distributed if $\underline{\underline{\phi_X(t) = \phi_Y(t)}} \; \forall t \in \mathbb{R}$.

Prerequisites

- Normal Random Variables
- Moment generating functions
- Characteristic functions
- Gamma function

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, dx$$

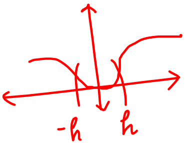
$$\Gamma(n) = (n-1)! \quad , n \in \mathbb{N}.$$

Prerequisites

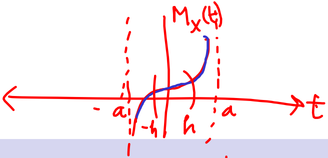
- Normal Random Variables
- Moment generating functions
- Characteristic functions
- Gamma function
- Power Series and Real Analyticness

$\to f^{(n)}(x)$ exists $\forall n$

$$0 \leftarrow e^{-1/x^2}, \quad x \neq 0$$
$$0, \quad x = 0$$



Before we begin



Theorem

$\forall k$ X, Y

If X and Y are two random variables such that $M_X^{(k)}(t)$ and $M_Y^{(k)}(t)$ exist, are equal to each other and are real analytic over some neighbourhood $(-h, h)$ containing 0, then the probability distributions of X and Y are the same. (1)

Before we begin

$Y = f(X_1, X_2, \ldots)$

$B(n,p) \sim \sum_{i=0}^{n} \text{Bernoulli}_i(p)$

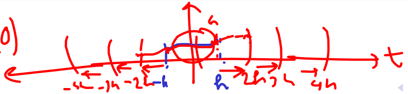
Theorem

If X and Y are two random variables such that $M_X^{(k)}(t)$ and $M_Y^{(k)}(t)$ exist, are equal to each other and are real analytic over some neighbourhood $(-h, h)$ containing 0, then the probability distributions of X and Y are the same. (1)

Proof.

A (complicated) proof can be found in Billingsley's *Probability and Measure*, sec. 30.1, pp. 388-89, but the gist of the proof is that you can express the characteristic function as $\varphi(t+x) = \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(t)}{k!} x^k$, $|x| < h$. Then taking $t = 0, h-\epsilon, -h+\epsilon, 2h-\epsilon, -2h+\epsilon \cdots$ causes the characteristic functions of X and Y to agree in $(-h, h)$, $(-2h, 2h)$, $(-3h, 3h)$ and so on. Hence, X and Y have the same distribution. □

$\varphi_X^{(k)}(0) = (i)^k M_X^{(k)}(0)$



Simple results on Normal Random Variables

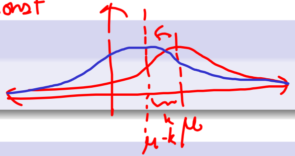
Theorem

If $X \sim N(\mu, \sigma^2)$, *then* $bX \sim N(b\mu, b^2\sigma^2)$

Simple results on Normal Random Variables

$X \sim N(\mu, \sigma^2)$; $X - k \sim N(\mu - k, \sigma^2)$

const



Theorem

If $X \sim N(\mu, \sigma^2)$, then $bX \sim N(b\mu, b^2\sigma^2)$

Proof.

Simple transformation of variables;

$f(h(x)) \left| \frac{d h(x)}{dx} \right|$

$$f_{bX}(x) = \frac{1}{b} f_X\left(\frac{x}{b}\right)$$

$$f_{bX}(x) = \frac{1}{\sqrt{2\pi} b\sigma} \, e^{\frac{(x-b\mu)^2}{2b^2\sigma^2}}$$

which is normally distributed with mean $b\mu$ and variance $b^2\sigma^2$ □

Linearity of Normal Random Variables I

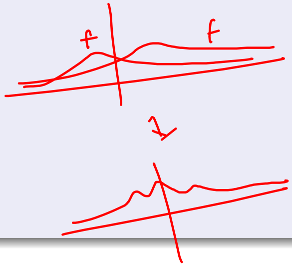
Theorem

If $X_1, \cdots, X_n$ is a sequence of independent random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$, then any linear combination

$$Y = \sum_{i=1}^{n} b_i X_i$$

is also a normal random variable. In particular,

$$Y \sim N\left(\sum_{i=1}^{n} b_i \mu_i, \ \sum_{i=1}^{n} b_i^2 \sigma_i^2\right)$$



Linearity of Normal Random Variables II

Proof.

Let $\mu = \sum_{i=1}^{n} b_i\mu_i$ and $\sigma^2 = \sum_{i=1}^{n} b_i^2\sigma_i^2$ for simplicity. Proceed by comparing the MGF's of Y and $N(\mu, \sigma^2)$.

$$
\begin{aligned}
M_Y(t) &= E(e^{t(b_1X_1+\cdots+b_nX_n)}) \\
&= \prod_{i=1}^{n} E(e^{tb_iX_i}) \\
&= \prod_{i=1}^{n} e^{b_i\mu_i t + b_i^2\sigma_i^2 t^2/2} \\
&= e^{t\overbrace{\sum_{i=1}^{n} b_i\mu_i}^{\mu} + \frac{t^2}{2}\overbrace{\sum_{i=1}^{n} b_i^2\sigma_i^2}^{\sigma^2}} \\
M_Y(t) &= e^{t\mu + t^2\sigma^2/2} \\
M_Y(t) &= M_{N(\mu,\sigma^2)}(t)
\end{aligned}
$$

$E(XY) = E(X)E(Y)$

X, Y ind.

And by the uniqueness of MGF, Y and $N(\mu, \sigma^2)$ are identically distributed. □

The χ^2 distribution

Definition

If a Random Variable X has the PDF

$$f_X(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\,\Gamma(k/2)}$$

Gamma distrib.

for nonnegative x and 0 otherwise, then X is said to follow a χ_k^2 distribution with k degrees of freedom.

The χ^2 distribution

Definition

If a Random Variable X has the PDF

$$f_X(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\,\Gamma(k/2)}$$

for nonnegative x and 0 otherwise, then X is said to follow a χ_k^2 distribution with k degrees of freedom.

this is not very useful in and of itself, so we'll explore the relationship between the ~~exponential~~ normal and the χ^2 distributions.

Relationship between χ^2 and normal random variables

$X_1^2 + X_2^2 \sim \chi_2^2$

Theorem

If $X_1, \cdots, X_n$ is a sequence of n iid random variables such that $X_i \sim \underline{N(0,1)}$, then $Y = \underbrace{\sum_{i=1}^{n} X_i^2}$ has a χ_n^2 distribution with n degrees of freedom

Relationship between χ^2 and normal random variables

Theorem

If $X_1, \cdots, X_n$ is a sequence of n iid random variables such that $X_i \sim N(0,1)$, then $Y = \sum_{i=1}^{n} X_i^2$ has a χ_n^2 distribution with n degrees of freedom

n

Proof.

Proceed by comparing the moment generating functions of the χ_n^2 distribution and the sum of n iid standard normal random variables.

$$-\left(\frac{1}{2}-t\right)$$

$$\left(\frac{1}{2}-t\right)x = x'$$

$$M_{\chi_n^2}(t) = \int_0^\infty \frac{x^{n/2-1} e^{\left(t-\frac{1}{2}\right)x}}{2^{n/2}\,\Gamma(n/2)}\,dx$$

$$= \frac{1}{\left(\frac{1}{2}-t\right)^{n/2}} \int_0^\infty \frac{\left(\left(\frac{1}{2}-t\right)x\right)^{n/2-1} e^{\left(t-\frac{1}{2}\right)x}}{2^{n/2}\,\Gamma(n/2)} \left(\frac{1}{2}-t\right) dx$$

$$\Gamma(n/2)$$

$$\boxed{M_{\chi_n^2}(t) = \frac{1}{(1-2t)^{n/2}}}$$

Relationship between χ^2 and normal random variables

Theorem

If $X_1, \cdots, X_n$ is a sequence of n iid random variables such that $X_i \sim N(0,1)$, then $Y = \sum_{i=1}^{n} X_i^2$ has a χ_n^2 distribution with n degrees of freedom

Proof.

$(X_1^2 + X_2^2 + \cdots + X_n^2)$

$$
\begin{aligned}
M_Y(t) &= E(e^{tY}) \\
&= \prod_{i=1}^{n} E(e^{tX_i^2}) \\
&= (M_{X_1^2}(t))^n
\end{aligned}
$$

$h(x) = \sqrt{x}$,

The PDF of X_1^2 is given by

$$
f_{X_1^2}(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}
$$

$f(h(x)) \left| \frac{dh(x)}{dx} \right|$

Relationship between χ^2 and normal random variables

Theorem

If $X_1, \cdots, X_n$ is a sequence of n iid random variables such that $X_i \sim N(0,1)$, then $Y = \sum_{i=1}^{n} X_i^2$ has a χ_n^2 distribution with n degrees of freedom

Proof.

Hence, the MGF of X_1^2 is

$$M_{X_1^2}(t) = \int_0^\infty \frac{x^{\frac{1}{2}-1} e^{\left(t-\frac{1}{2}\right)x}}{\sqrt{2\pi}}\,dx$$

$$= \frac{1}{\left(\frac{1}{2}-t\right)^{1/2}} \int_0^\infty \frac{\left(\left(\frac{1}{2}-t\right)x\right)^{1/2-1} e^{\left(t-\frac{1}{2}\right)x}}{\sqrt{2\pi}} \left(\frac{1}{2}-t\right) dx$$

$$\left(\frac{1}{2}-t\right)x = x'$$

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

$$\underline{\underline{M_{X_1^2}(t)}} = \frac{1}{\sqrt{1-2t}}$$

$$\frac{1}{(1-2t)^{n/2}} = M_{\chi_n^2}(t)$$

Relationship between χ^2 and normal random variables

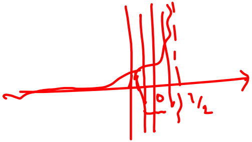
Theorem

If $X_1, \cdots, X_n$ is a sequence of n iid random variables such that $X_i \sim N(0,1)$, then $Y = \sum_{i=1}^{n} X_i^2$ has a χ_n^2 distribution with n degrees of freedom

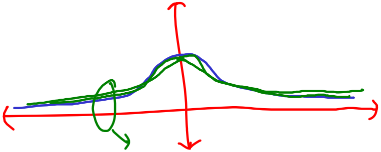
Proof.

This implies that the MGF of Y is $\frac{1}{(1-2t)^{n/2}}$, which is the same as χ_n^2. Equivalence of distributions of Y and χ_n^2 follows from (1). Note that the MGF exists only when $|t| < 1/2$, which, as proven by (1), is a sufficient condition for the distributions to be the same. $\square$

$1-2t < 0$



The t distribution



Definition

If a random variable X has the probability distribution given by

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

then X is said to follow the t-distribution with ν degrees of freedom.

$\to \infty$

$$\nu \to \infty, \quad t \xrightarrow{P} \underline{\underline{N(0,1)}}$$

The t distribution

Definition

If a random variable X has the probability distribution given by

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

then X is said to follow the t-distribution with ν degrees of freedom.

Again, this gives us no intuition about the t distribution. To motivate the distribution, we will need to explore sampling first.

Sampling

?

$f(x;\theta)$

8R 6G 8B 100%

statistical inference

fair/not.

$n(G) = 2n(R) = 2n(B)$?

≈

? Yes but no?

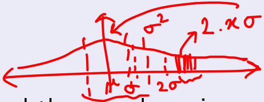
Definition

If the random variables $X_1, \ldots, X_n$ are independent and identically distributed, then these random variables constitute a random sample of size n from the common distribution, which has a mean μ and a ~~standard deviation~~ Variance σ^2

≠ normal.

$X_i \sim N(\mu, \sigma^2)$

We define the sample mean as



$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow N(\mu, \sigma^2)$$

?

$$\frac{X_i}{n} \sim N\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right)$$

and the sample variance as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

bessel's correction

N

$$\boxed{\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}}$$

$$\boxed{\sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

The denominator here is $n-1$ instead of n so that $E(S^2) = \sigma^2$, that is, to ensure that the sample variance is unbiased.

Z scores and Z tests

Definition

The Standard score (or Z score) of a sample X is the standardized ratio of the sample mean from the mean

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

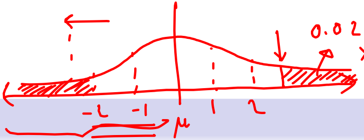
This can be used to estimate how far off X is from the mean if the mean is known, or give error bounds for the mean if the mean is unknown. ELL101 flashbacks for some of you

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim N(0, 1)$$

Z scores and Z tests



0.0297 < 0.05

×2

0.0594 > 0.05

milk

w ~ 40%

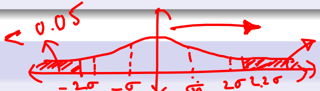
w ~ 20%

Definition

The Standard score (or Z score) of a sample X is the standardized ratio of the sample mean from the mean

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

This can be used to estimate how far off X is from the mean if the mean is known, or give error bounds for the mean if the mean is unknown. ELL101 flashbacks for some of you



< 0.05

milk dilution

Definition

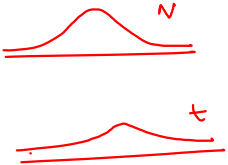
The Z test is used to see how far off a sample is from the mean, assuming that the distribution of the population is standard normal (this is assumed to be true given a large population because of the central limit theorem). For a one-tailed test, $\Phi(Z)$ or $1 - \Phi(Z)$ gives the probability of the event that a given sample is picked.

The t score

$100 \to \bar{X}, S^2 \to \mu, \sigma^2$

Note that the Z test only works when the sample mean and variance are known, or can be approximated easily. For small samples (of size 5-10), the variance cannot be approximated easily, and comparatively 'larger' error bounds are needed if the mean is to be calculated. To account for this, we simply replace the standard deviation σ with the sample standard deviation S, and the corresponding random variable

$$T = \underbrace{\frac{\overline{X} - \mu}{S/\sqrt{n}}}_{?\,\text{Normal } X}$$



is called the T score.

The t distribution and the t test

Definition

The Probability density function of the random variable $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ is a t distribution with $n-1$ degrees of freedom.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N, \quad \frac{S^2}{\sigma^2} \rightarrow \chi^2_{n-1} \quad \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n}}} \sim t$$

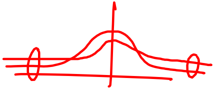
The t distribution and the t test

Definition

The Probability density function of the random variable $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ is a t distribution with $n-1$ degrees of freedom.

Definition

The t test is similar to the Z test, but rather than use the CDF of the standard normal, we now use the CDF of the t distribution, because the t score obeys a t distribution. The t distribution is more 'heavy-tailed', with the distribution converging to the standard normal as $n \to \infty$.



References

- Billingsley, Patrick. *Probability and Measure*. 3rd ed, Wiley, 1995.
- Hogg, Robert V., et al. *Introduction to Mathematical Statistics*. 7th ed, Pearson, 2013.
- Scholz, Fritz. *Elements of Statistical methods: slides from Stat311 course at University of Washington*. 2010.
- Student. "The Probable Error of a Mean." *Biometrika*, vol. 6, no. 1, Mar. 1908, p. 1. DOI.org (Crossref), `https://doi.org/10.2307/2331554`. (This is the original paper by Gosset, a readable copy can be found here)
- Too many Wikipedia articles