Lecture 14

- Moments
- KDE
- Statistical Inference
- Hypothesis tests

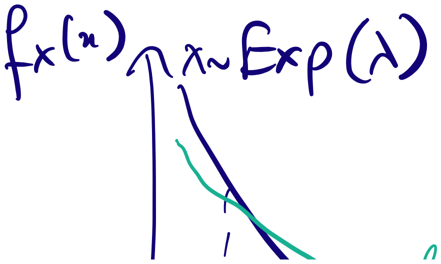
---

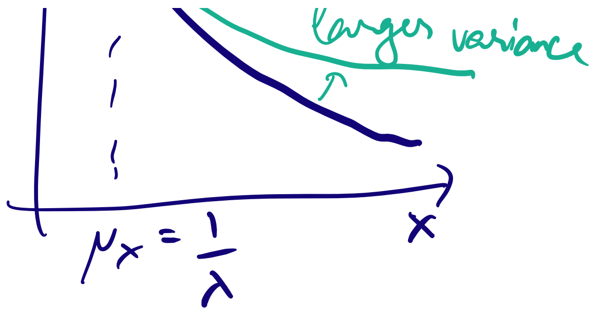
n^{th} central moment (discrete RV):

$$E[(X-\mu_X)^n] = \sum_x (x-\mu_X)^n \cdot p_X(x)$$

1st central moment = mean

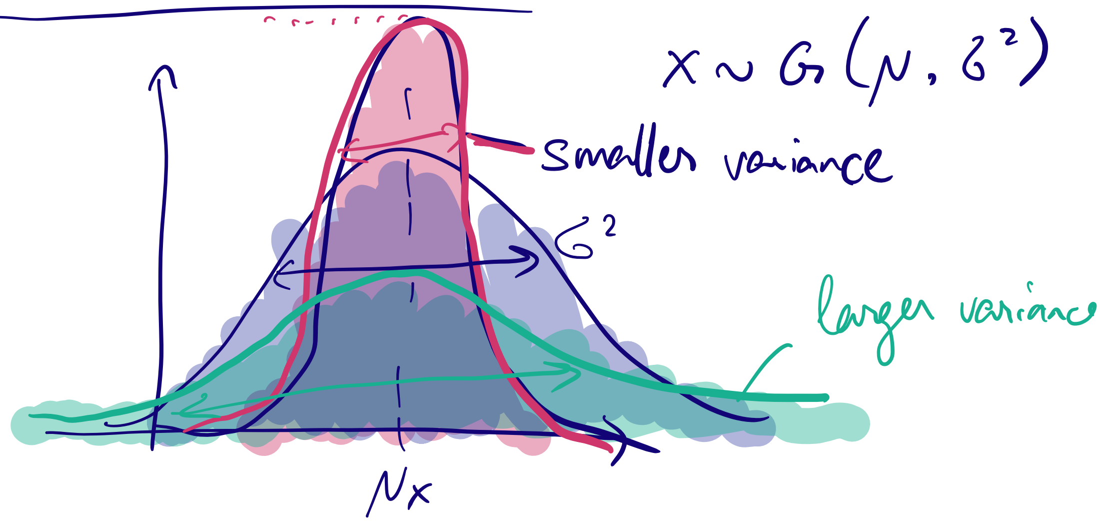
$$\boxed{E[X] = \mu_X \neq \text{mean}}$$





2^{nd} Central Moment $\equiv Var[x]$

$x \sim G(\mu, \sigma^2)$



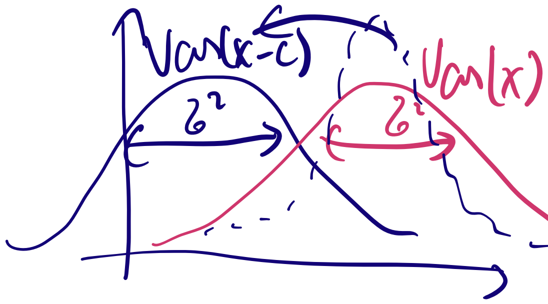
$$Var[x] = E[(x-\mu_x)^2], \quad \text{where } \mu_x = E[x]$$

$$= E[x^2 - 2\mu_x x + \mu_x^2]$$

$$= E[x^2] - 2\mu_x \underbrace{E[x]}_{=\mu_x} + \mu_x^2$$

$$= E[x^2] - \mu_x^2$$

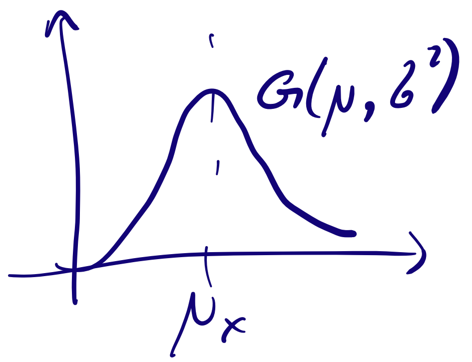
$$= E[x^2] - (E[x])^2$$



$$Var(x-c) = Var(x)$$

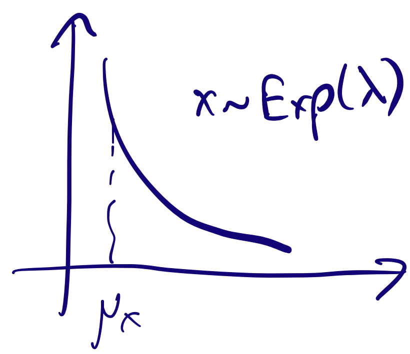
3rd central moments = skewness

$$E[(X-\mu_X)^3]$$



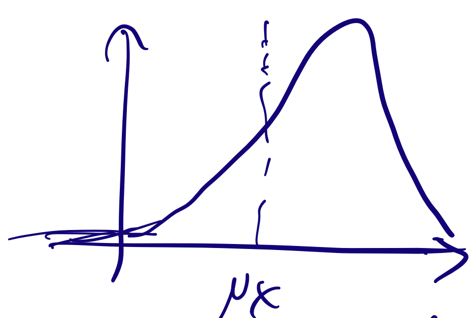
symmetric w.r.t μ_X

skewness $= 0$



long right tail

skewness > 0



$X \sim Binomial(15, 0.9)$

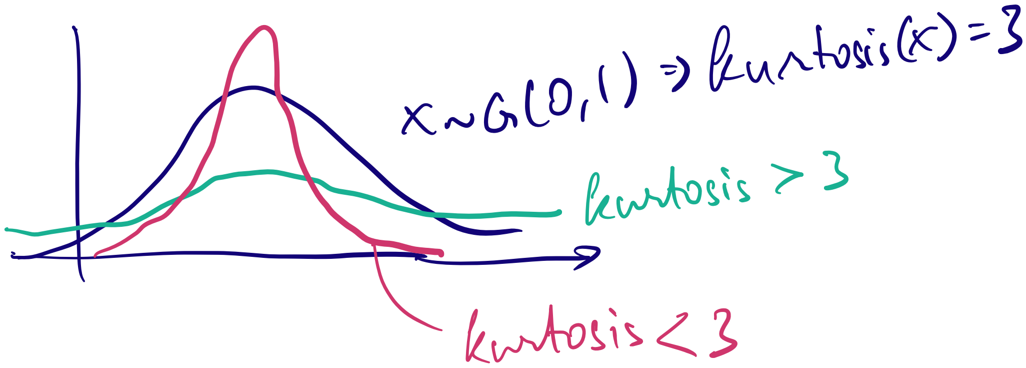
Long left tail

Skewness < 0

4^{th} central moment $\equiv$ kurtosis

$E[(X-\mu_X)^4]$

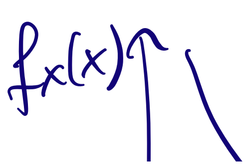
measures the "volume" of the tails in the distribution
(→ area under the curve)



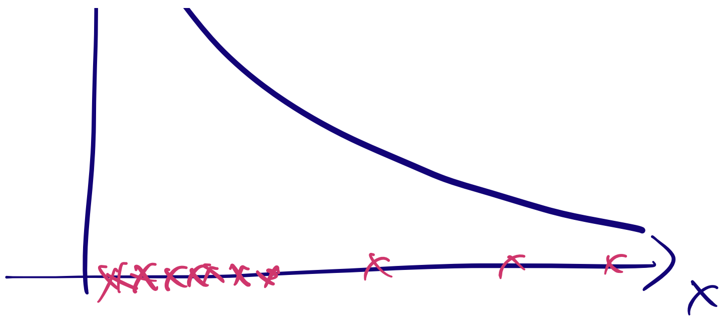
$E[(X-\mu_X)^4] - 3 \equiv$ "Excess" kurtosis

$= 0$ for $G(0,1)$

Statistical Inference



$X \sim \text{Exp}(\lambda)$



λ is a <u>parameter</u> that we can <u>infer</u> from data!

θ = parameters ← deterministic but unknown!

$f_X(x;\theta)$ = PDF with unknown parameters θ

If $x \sim Exp(\lambda)$, then $\theta = \lambda$

If $x \sim G(\mu, \sigma^2)$, then $\theta = \{\mu, \sigma^2\}$

$\hat{\theta}$ = estimate

$\hat{\theta}$ is estimated from a data sample

$$\{x_i\}_{i=1}^{N} = \{x_1, \ldots, x_N\}$$

Error of that estimator:

$$\mathcal{E}_\theta(\hat{\theta}) = \hat{\theta} - \theta$$

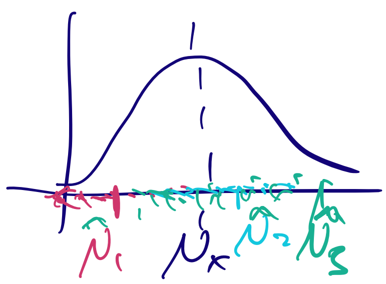
Bias of estimator

$$b_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta$$

If $E[\hat{\theta}] = \theta$
then $b_\theta(\hat{\theta}) = 0$
$\Downarrow$
unbiased estimator!

Variance of estimator

$$Var_\theta[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$$



$$E[\hat{\mu}] = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i \rightarrow \mu_x$$

Mean-Square error (MSE)

$$E[(\theta - \hat{\theta})^2] = b_\theta^2(\hat{\theta}) + Var_\theta(\hat{\theta})$$

Data samples

$$X = \{x_1, \dots x_N\} = \{x_i\}_{i=1}^{N}$$

sample size $= N$

Draw from the same underlying distribution independently:

independent & identically distributed (i.i.d.)

Estimator for the mean:

True mean $= \mu_x$

Estimate $= \hat{\mu}_x$

$$\boxed{\hat{\mu}_x = \frac{1}{N} \sum_{i=1}^{N} x_i} \Rightarrow \hat{\mu}_x \text{ is an unbiased estimator}$$

of μ_x

Proof:

$$E[\hat{\mu}_x] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N} E\left[\sum_{i=1}^{N} x_i\right]$$

$$= \frac{1}{N} E[x_1 + x_2 + \dots + x_N]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \underbrace{E[x_i]}_{=\mu_x} = \frac{1}{N}\sum_{i=1}^{N}\mu_x = \mu_x$$

Estimator for the variance:

True variance is σ_x^2

(1) $$\boxed{s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2}$$

1st estimator for variance

variance for R.V. x:

$$\sigma_x^2 = E[(x-\mu_x)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)^2$$

$$\boxed{E[s_N^2] = \frac{N-1}{N}\cdot\sigma_x^2} \neq \sigma_x^2$$

biased estimator!

Note that $\frac{N-1}{N} \xrightarrow{N\to\infty} 1$

$$\frac{\quad}{N} \longrightarrow 1$$

$\Downarrow$

Sample size N is large, then we may get unbiased estimator

② $$\frac{N}{N-1} s_N^2 = \boxed{S_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2}$$

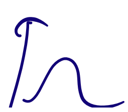
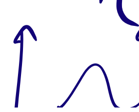
2nd estimator for variance

$$E[S_{N-1}^2] = \sigma_x^2 \Rightarrow$$ unbiased estimator!

Properties of sum of independent Gaussian R.V.s

$X = \{x_i\}_{i=1}^{N}$ $\qquad$ $Y = \{y_i\}_{i=1}^{N}$

$X \sim G(\mu_x, \sigma_x^2)$ $\qquad$ $Y \sim G(\mu_y, \sigma_y^2)$

$$Z = X + Y$$

If x & y are independent RVs:

$$\mu_Z = \mu_X + \mu_Y$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

$$Z \sim G(\mu_Z, \sigma_Z^2)$$

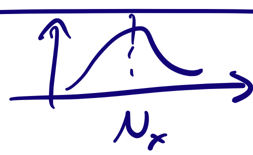
$Z' = aZ + b$ is a Gaussian R.V.

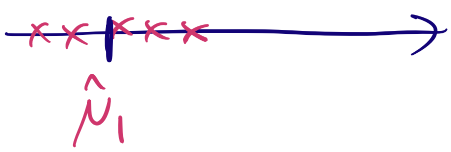
$$\mu_{Z'} = a\mu_Z + b$$

$$\sigma_{Z'}^2 = a^2 \sigma_Z^2$$

---

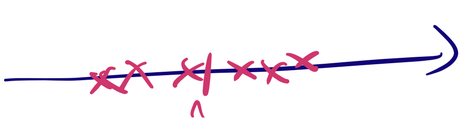
① sample 1





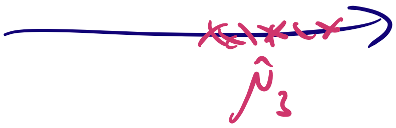
$$\hat{\mu}_i = \frac{1}{N} \sum_{j=1}^{N} x_j$$

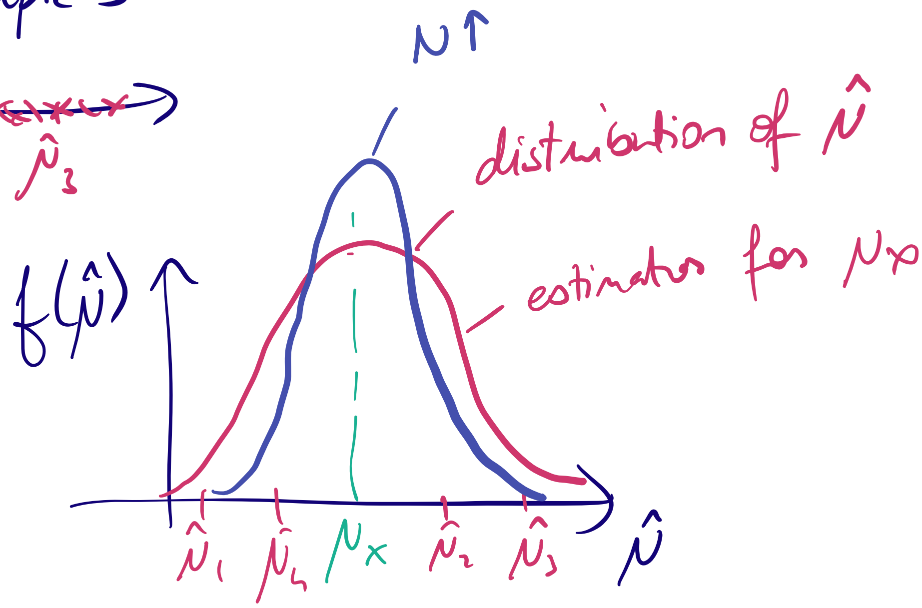
② sample 2



μ_2

③ sample 3





$$\hat{\mu}_x \sim G\left(\mu_x, \frac{\sigma_x^2}{N}\right)$$

$$\begin{aligned}
Var[\hat{\mu}_x] &= var\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] \\
&= \frac{1}{N^2} var\left[\sum_{i=1}^{N} x_i\right] \\
&= \frac{1}{N^2} \cdot var[x_1 + \dots + x_N] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \underbrace{var[x_i]}_{\sigma_x^2}
\end{aligned}$$

var(a·x)
$= a^2 var(x)$
for a constant

$$= \frac{1}{N^2} \cdot N \cdot \sigma_x^2$$

$$= \frac{1}{N} \cdot \sigma_x^2 \xrightarrow[N \to \infty]{} 0$$

Hypothesis Testing:

z-test

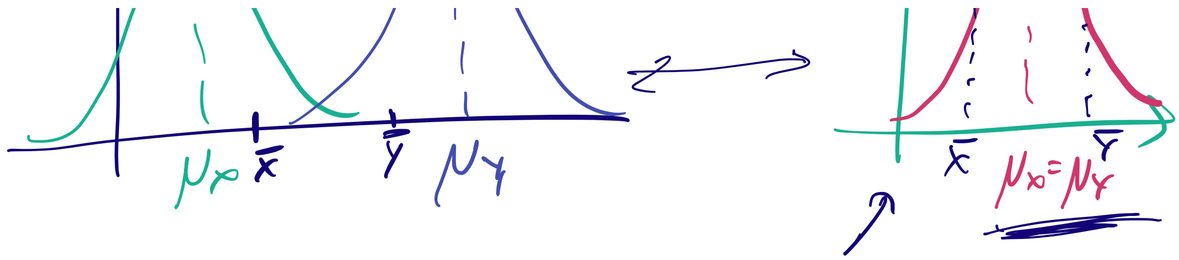
$X = \{x_i\}_{i=1}^{N}$ $\qquad$ $Y = \{y_i\}_{i=1}^{M}$

N samples for x $\qquad$ M samples for Y

Consider the estimator for the mean of these two sampled data:

$$\boxed{\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i} \quad \text{and} \quad \boxed{\bar{y} = \frac{1}{M} \sum_{j=1}^{M} y_j}$$

If $\bar{x} \neq \bar{y} \Rightarrow \mu_x \neq \mu_y$



z-test checks if two sampled data are drawn from the same distribution by measuring the difference of their mean estimator

It assumes that the variance for the mean estimators are known:

$$Var[X] = var[Y] = \sigma^2$$

$$\boxed{T = \hat{\mu}_x - \hat{\mu}_y}$$ ← this is also Gaussian distributed

$$t = \bar{x} - \bar{y}$$

under the assumption that $\mu_x = \mu_Y = \mu$

① $E[T] = \mu_T = E[\hat{\mu}_x - \hat{\mu}_Y]$

$$= E[\hat{\mu}_x] - E[\hat{\mu}_Y]$$

$$= \mu - \mu$$
$$= 0$$

② $\sigma_T^2 = \mathrm{var}[T]$

$$= \mathrm{var}[\hat{\mu}_X - \hat{\mu}_Y]$$
$$= \mathrm{var}[\hat{\mu}_X + (-1)\cdot\hat{\mu}_Y] \quad \text{add}$$
$$= \mathrm{var}[\hat{\mu}_X] + \mathrm{var}[(-1)\cdot\hat{\mu}_Y] \quad \mathrm{var}(cx) = c^2\,\mathrm{var}(x)$$
$$= \mathrm{var}[\hat{\mu}_X] + \mathrm{var}[\hat{\mu}_Y]$$
$$= \frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M} = \sigma^2\left(\frac{1}{N} + \frac{1}{M}\right)$$

assumption: known variances for x & Y

(z-test)

assumption: equal variances for x and Y

Hypothesis Test :

H_0 : the true means are the same

$$\mu_X = \mu_Y$$

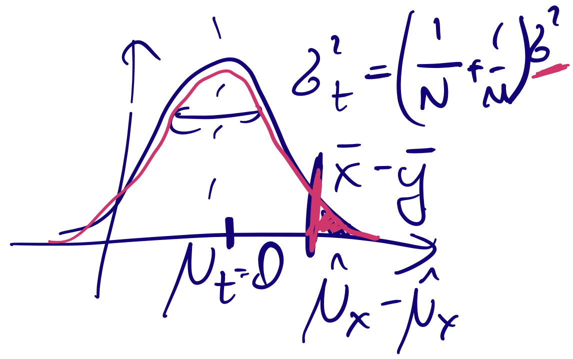
H_1: $\mu_X \neq \mu_Y$

Steps:

① compute the statistic : $t = \hat{\mu}_X - \hat{\mu}_Y$

② we know that t is Gaussian distributed ; we know the mean & variance of this distribution under the null hypothesis.

③ compute p-value



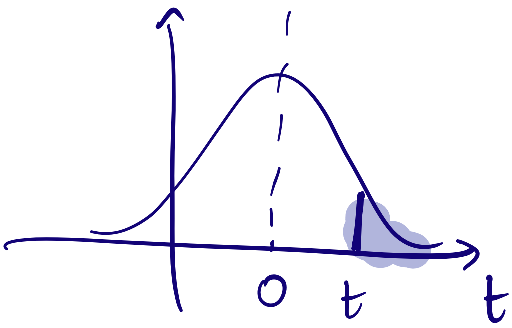
H_0 :

$$T = \hat{\mu}_X - \hat{\mu}_Y$$

$$T \sim G\left(0,\ \sigma^2\left(\frac{1}{N} + \frac{1}{M}\right)\right)$$

$$P(T \geq t \mid H_0) = P(T > t \mid H_0)$$

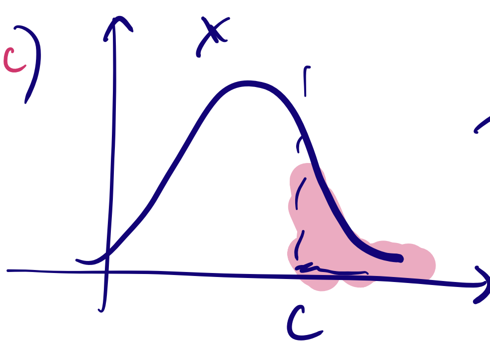
$$= Q\left(\frac{t - \mu_T}{\sigma_T}\right)$$



$$= Q\left(\frac{t}{\sqrt{\sigma^2\left(\frac{1}{N} + \frac{1}{M}\right)}}\right)$$

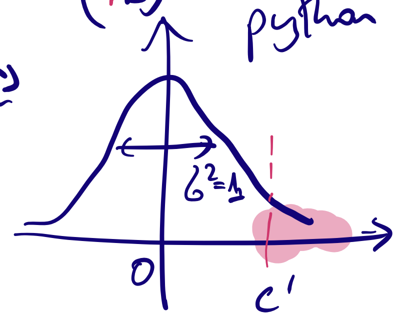
SA3

(1c)



transform

(1b) python



(1a) Q-function table