# Mathematical statistics 

September 6 ${ }^{\text {th }}, 2018$
Lecture 4: The distribution of a linear combination

| Week 1 | Probability reviews |
| :---: | :---: |
|  | Chapter 6: Statistics and Sampling Distributions |
| Week 4 | Chapter 7: Point Estimation |
| Week 7 | Chapter 8: Confidence Intervals |
| Week 10 | Chapter 9: Test of Hypothesis |
| Week 14 | Regression |

## Overview

6.1 Statistics and their distributions
6.2 The distribution of the sample mean
6.3 The distribution of a linear combination

Order $6.1 \rightarrow 6.3 \rightarrow 6.2$

## Random sample



## Definition

The random variables $X_{1}, X_{2}, \ldots, X_{n}$ are said to form a (simple) random sample of size $n$ if
(1) the $X_{i}$ 's are independent random variables
(2) every $X_{i}$ has the same probability distribution

## Definition

A statistic is any quantity whose value can be calculated from sample data

- prior to obtaining data, there is uncertainty as to what value of any particular statistic will result $\rightarrow$ a statistic is a random variable
- the probability distribution of a statistic is referred to as its sampling distribution


## Example of a statistic

- Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample of size $n$
- The sample mean of $X_{1}, X_{2}, \ldots, X_{n}$, defined by

$$
\bar{X}=\frac{X_{1}+X_{2}+\ldots X_{n}}{n}
$$

is a statistic

- When the values of $x_{1}, x_{2}, \ldots, x_{n}$ are collected,

$$
\bar{x}=\frac{x_{1}+x_{2}+\ldots x_{n}}{n}
$$

is a realization of the statistic $\bar{X}$

## Example of a statistic

- Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample of size $n$
- The random variable

$$
T=X_{1}+2 X_{2}^{2}+3 X_{5}^{3}
$$

is a statistic

- When the values of $x_{1}, x_{2}, \ldots, x_{n}$ are collected,

$$
t=x_{1}+2 x_{2}^{2}+3 x_{5}^{3},
$$

is a realization of the statistic $T$

## Questions for this chapter

Given a random sample $X_{1}, X_{2}, \ldots, X_{n}$, and

$$
T=a_{1} X_{1}+a_{2} X_{2}+\ldots+a_{n} X_{n}
$$

- If we know the distribution of $X_{i}$ 's, can we obtain the distribution of $T$ ?
- Last lecture: Simple cases
- $2^{\text {nd }}$ lecture: If $X_{i}^{\prime} s$ follow normal distribution, then so does $T$.
- If we don't know the distribution of $X_{i}$ 's, can we still obtain/approximate the distribution of $T$ ?
- Can we at least compute the mean and the variance?
- When $T$ is the sample mean, i.e. $a_{1}=a_{2}=\ldots=\frac{1}{n}$


## Last lecture

If $T=X_{1}+X_{2}$

- compute the distribution of $T$ in some easy cases
- compute the expected value and variance of $T$


## Example 1

## Problem

Consider the distribution $P$

| $x$ | 10 | 15 | 20 |
| :---: | :---: | :---: | :---: |
| $p(x)$ | 0.2 | 0.3 | 0.5 |

Let $\left\{X_{1}, X_{2}\right\}$ be a random sample of size 2 from $P$, and $T=X_{1}+X_{2}$.
(1) Compute $P[T=100]$
(2) Derive the probability mass function of $T$
(3) Compute the expected value and the standard deviation of $T$

## Example 2

## Problem

Let $\left\{X_{1}, X_{2}\right\}$ be a random sample of size 2 from the exponential distribution with parameter $\lambda$

$$
f(x)= \begin{cases}\lambda e^{-\lambda x} & \text { if } x \geq 0 \\ 0 & \text { if } x<0\end{cases}
$$

and $T=X_{1}+X_{2}$.
What is the distribution of $T$ ?
(1) If the distribution and the statistic $T$ is simple, try to construct the pmf of the statistic (as in Example 1)
(2) If the probability density function $f_{X}(x)$ of $X$ 's is known, the

- try to represent/compute the cumulative distribution (cdf) of $T$

$$
\mathbb{P}[T \leq t]
$$

- take the derivative of the function (with respect to $t$ )


## Linear combination of normal random variables


$E(X)=\mu, \operatorname{Var}(X)=\sigma^{2}$

## Shaded area $=\Phi(z)$



Table A. 3 Standard Normal Curve Areas (cont.) $\quad \Phi(z)=P(Z \leq z)$

| $\boldsymbol{z}$ | $\mathbf{. 0 0}$ | $\mathbf{. 0 1}$ | $\mathbf{. 0 2}$ | $\mathbf{. 0 3}$ | $\mathbf{. 0 4}$ | $\mathbf{. 0 5}$ | $\mathbf{. 0 6}$ | $\mathbf{. 0 7}$ | $\mathbf{. 0 8}$ | $\mathbf{. 0 9}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9278 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |

## Shifting and scaling normal random variables

## Problem

Let $X$ be a normal random variable with mean $\mu$ and standard deviation $\sigma$.
Then

$$
Z=\frac{X-\mu}{\sigma}
$$

follows the standard normal distribution.

## Shifting and scaling normal random variables

If $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, then

$$
Z=\frac{X-\mu}{\sigma}
$$

has a standard normal distribution. Thus

$$
\begin{gathered}
P(a \leq X \leq b)=P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\
=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right) \\
P(X \leq a)=\Phi\left(\frac{a-\mu}{\sigma}\right) \quad P(X \geq b)=1-\Phi\left(\frac{b-\mu}{\sigma}\right)
\end{gathered}
$$

## Linear combination of normal random variables

## Theorem

Let $X_{1}, X_{2}, \ldots, X_{n}$ be independent normal random variables (with possibly different means and/or variances). Then

$$
T=a_{1} X_{1}+a_{2} X_{2}+\ldots+a_{n} X_{n}
$$

also follows the normal distribution.
What are the mean and the standard deviation of $T$ ?

- $E(T)=a_{1} E\left(X_{1}\right)+a_{2} E\left(X_{2}\right)+\ldots+a_{n} E\left(X_{n}\right)$
- $\sigma_{T}^{2}=a_{1}^{2} \sigma_{X_{1}}^{2}+a_{2}^{2} \sigma_{X_{2}}^{2}+\ldots+a_{n}^{2} \sigma_{X_{n}}^{2}$


## Example 1

## Problem

Assume that

$$
X_{1} \sim \mathcal{N}(50,13) \quad \text { and } \quad X_{2} \sim \mathcal{N}(30,12)
$$

are independent.

- What is the distribution of $T=X_{1}-X_{2}$ ?
- What is $P[T \leq 29.8]$


## Example 2

## Problem

A concert has three pieces of music to be played before intermission. The time taken to play each piece has a normal distribution.
Assume that the three times are independent of each other. The mean times are 15,30 , and 20 min , respectively, and the standard deviations are 1,2 , and 1.5 min , respectively.

What is the distribution of the length of the concert?

## Example 3

## Problem

Let $X_{1}, X_{2}, \ldots, X_{16}$ be a random sample from $\mathcal{N}(1,4)$ (that is, normal distribution with mean $\mu=1$ and standard deviation $\sigma=2$ ).
Let $\bar{X}$ be the sample mean

$$
\bar{X}=\frac{X_{1}+X_{2}+\ldots+X_{16}}{16}
$$

- What is the distribution of $\bar{X}$ ?
- Compute $P[\bar{X} \leq 1.82]$


## Example 3*

## Problem

Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample from $\mathcal{N}\left(\mu, \sigma^{2}\right)$ (that is, normal distribution with mean $\mu$ and standard deviation $\sigma$ ). Let $\bar{X}$ be the sample mean

$$
\bar{X}=\frac{X_{1}+X_{2}+\ldots+X_{n}}{n}
$$

What is the distribution of $\bar{X}$ ?

## Example 4

## Problem

Two airplanes are flying in the same direction in adjacent parallel corridors. At time $t=0$, the first airplane is 10 km ahead of the second one.
Suppose the speed of the first plane ( $\mathrm{km} / \mathrm{h}$ ) is normally distributed with mean 520 and standard deviation 10 and the second planes speed, independent of the first, is also normally distributed with mean and standard deviation 500 and 10, respectively.

What is the probability that after $2 h$ of flying, the second plane has not caught up to the first plane?

## What if $X_{i}$ 's are not normal distributions?

## Linear combination of random variables

## Theorem

Let $X_{1}, X_{2}, \ldots, X_{n}$ be independent random variables (with possibly different means and/or variances). Define

$$
T=a_{1} X_{1}+a_{2} X_{2}+\ldots+a_{n} X_{n}
$$

then the mean and the standard deviation of $T$ can be computed by

- $E(T)=a_{1} E\left(X_{1}\right)+a_{2} E\left(X_{2}\right)+\ldots+a_{n} E\left(X_{n}\right)$
- $\sigma_{T}^{2}=a_{1}^{2} \sigma_{X_{1}}^{2}+a_{2}^{2} \sigma_{X_{2}}^{2}+\ldots+a_{n}^{2} \sigma_{X_{n}}^{2}$


## Mean and variance of the sample mean

## Problem

Given a random sample $X_{1}, X_{2}, \ldots, X_{n}$ from a distribution with mean $\mu$ and standard deviation $\sigma$, the mean is modeled by a random variable $\bar{X}$,

$$
\bar{X}=\frac{X_{1}+X_{2}+\ldots+X_{n}}{n}
$$

- Compute $E(\bar{X})$
- Compute $\operatorname{Var}(\bar{X})$


## Mean and variance of the sample mean

Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $E(\bar{X})=\mu_{\bar{X}}=\mu$
2. $V(\bar{X})=\sigma_{\bar{X}}^{2}=\sigma^{2} / n$ and $\sigma_{\bar{X}}=\sigma / \sqrt{n}$

## Theorem

Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^{2}$. Then, in the limit when $n \rightarrow \infty$, the standardized version of $\bar{X}$ have the standard normal distribution

$$
\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma / \sqrt{n}} \leq z\right)=\mathbb{P}[Z \leq z]=\Phi(z)
$$

Rule of Thumb:
If $n>30$, the Central Limit Theorem can be used for computation.

## Example: population distribution



Matt Nedrick (2015).
http://github.com/mattnedrich/CentralLimitTheoremDemo

## Sample distribution: $n=3$



## Sample distribution: $n=10$



## Sample distribution: $n=30$



## Example 4

## Problem

When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g .

If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity $X$ is between 3.5 and 3.8 g ?

Hint:

- First, compute $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$
- Note that

$$
\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}}
$$

is (approximately) standard normal.

