<<Lecture5_SSNotes.pdf>>

Stat 324 – Introduction to Statistics for Engineers

LECTURE 5: TRANSFORMING AND COMBINING RANDOM VARIABLES AND SAMPLING DISTRIBUTIONS (OL:4.12, 4.14)

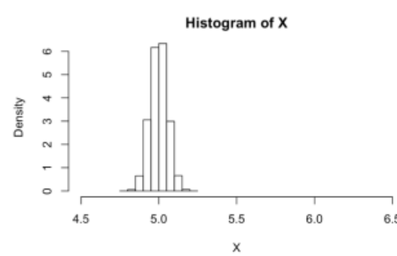
Transformations of Random Variables

In practice, we often construct new random variables by performing arithmetic operations on other random variables. We would like to be able to describe the resulting distributions.

Suppose we know the distribution of

X: length of steel rod produced by certain machine

E(X): $\mu_X = 5\ in$, variance(X): $\sigma_X^2 = 0.003\ in^2$



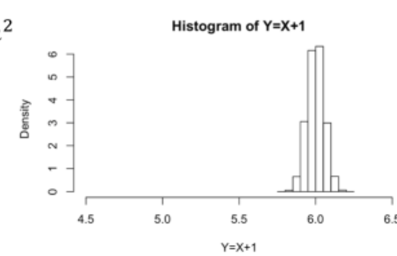
1. Adding a Constant: Suppose each rod is attached to a base that is exactly 1 in long

   Define Y: length of steel rod attached to base, Y=X+1

   E(Y): $\mu_{Y=X+1} =$ __6__, variance(X): $\sigma_{Y=X+1}^2 = \sigma_X^2 =$ 0.003 in^2

E(X+c)= E(x) + C

Var(X+c)= Var (x)



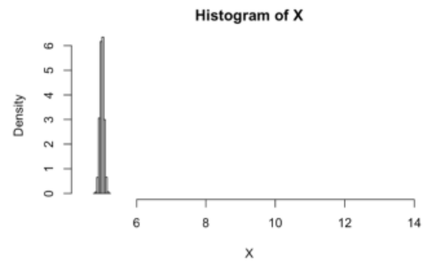
Transformations of Random Variables

In practice, we often construct new random variables by performing arithmetic operations on other random variables. We would like to be able to describe the resulting distributions.

Suppose we know the distribution of

X: length of steel rod produced by certain machine

E(X): $\mu_X = 5\ in$, variance(X): $\sigma_X^2 = 0.003\ in^2$



2. Multiplying by a Constant: Suppose each rod is measured in centimeters rather than inches.

Define Y: length of steel rod in centimeters, Y=2.54X

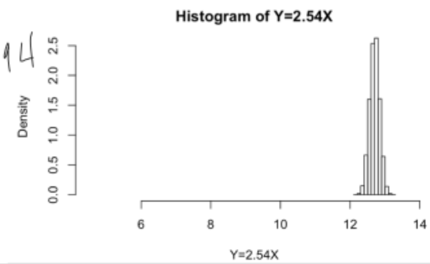
E(Y): $\mu_{Y=2.54X} = 2.54 * E(X) = 2.54 * 5 = 12.7$

variance(Y): $\sigma^2_{Y=2.54X} = 2.54^2 * Var(X) = 2.54^2 * 0.003 = 0.0194$

E(cX)= $C * E(X)$

Var(cX)= $c^2 * Var(X)$

SD(cX)= $|c| * SD(X)$



Transformations of Random Variables

3. Linear Transformation: Y=aX+b:

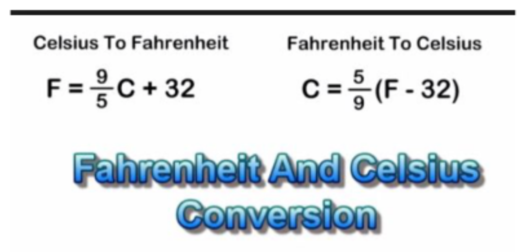
e.g.: Changing from Fahrenheit to/from Celsius

e.g.: Converting to Standardized Scores

E(aX+b)= $a * E(X) + b$

Var(aX+b)= $a^2 * Var(X)$

SD(aX+b)= $|a| * SD(X)$



Suppose $X \sim N(\mu, \sigma^2)$ *and* $Z = \frac{X-\mu}{\sigma}$

$E(Z) =$ $E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}(X-\mu)\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}\left[E(X)-\mu\right] = \frac{1}{\sigma}\left[\mu-\mu\right] = \frac{0}{\sigma} = 0$

$\mathrm{Var}(Z) =$ $var\left(\frac{1}{\sigma}(X-\mu)\right) = \frac{1}{\sigma^2}var(X-\mu) = \frac{1}{\sigma^2}var(X) = \frac{\sigma^2}{\sigma^2} = 1$

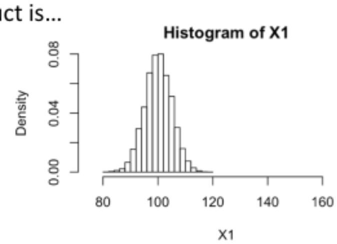
$Z = \frac{X-\mu}{\sigma} \sim$ *Non-Trivial to Prove Distribution is normal

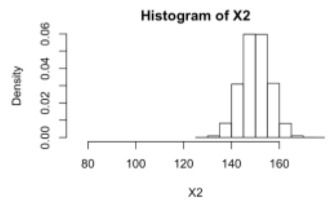
Suppose $X \sim N(\mu, \sigma^2)$, $Y = aX + b \sim N(a*\mu+b, a^2\sigma^2)$ Transformations of Normal Variables are also Normal

Combinations of Random Variables continued

Suppose we have multiple random variables: **X1, X2, X3,** …

E.g. Assume that two machines fabricate separate metal parts that are welded together to get one product. The mean length of part A is 100 mm with sd=5, and the mean length of part B is 150 mm with sd=6. The mean and sd length of the final combined product is…

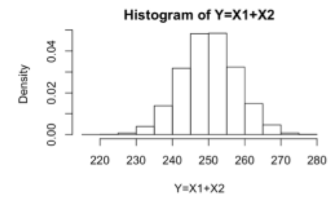




E(X1+X2+…)= $E(X_1) + E(X_2) \dots + E(X_n)$

E(aX1+bX2)= $a * E(X_1) + b * E(X_2)$

If X1 and X2 are **independent** random variables, then we can calculate variance easily:



Var(X1±X2)= $Var(X_1) + Var(X_2)$

Var(aX1±bX2)= $a^2 * Var(X_1) + b^2 * Var(X_2)$

$\mu_{X_1+X_2} = 250$

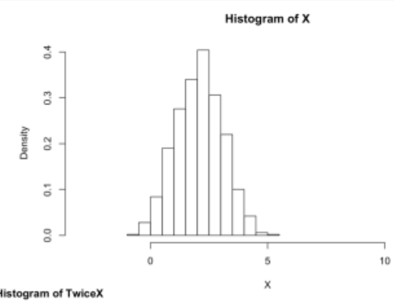
If we have multiple **Normally Distributed** random variables: **X1, X2, X3,**

Then, any **linear combination** ($\sum_{i=1}^{n} a_i X_i$**)** of the Xi is also Normally Distributed.

A Reminder about Notation Functionality

Suppose X is a random variable with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$

What is the difference between



Y1=2X

*Sampling once from X

*Multiplying value by 2

* $E(Y_1) = 2E(X) = 2\mu$

* $\mathrm{Var}(Y_1) = 2^2 \, Var(X) = 4\sigma^2$

* $\mathrm{SD}(Y_1) = 2\sigma$

and

Y2=X+X

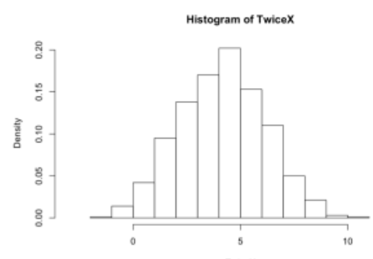
*Sampling twice from X

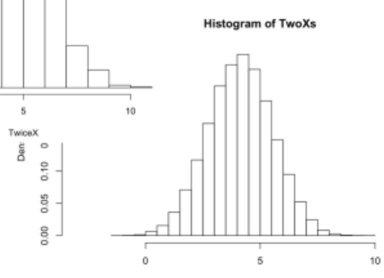
*add two realizations together

* $E(Y_2) = E(X_1) + E(X_2) = \mu + \mu = 2\mu$

intec independent * $\mathrm{Var}(Y_2) = Var(X_1) + Var(X_2) = 2\sigma^2$

* $\mathrm{SD}(Y_2) = \sqrt{2}\sigma$





Random Variables in Sampling

Reviewing a few concepts:

A sample from a population is called a **simple random sample** if every possible sample is _equally likely_ to be drawn. Unless otherwise specified, all samples in this class are SRS.

We say a sample is drawn **with replacement** if an element is _replaced_ to the population before the next element is drawn. Otherwise we say the sample is drawn **without replacement** (and every element can be drawn at most once.

A collection of RVs X1, X2, …Xn are said to be **independent and identically distributed (iid)** if:

1. The RVs are all _independent_ of each other
   - e.g the cholesterol score of 2 randomly chosen people
   - non e.g. the cholesterol reading of _1_ person in January and February

2. They all have exactly the same probability distribution.

We want to think of RVs as populations and a sample as realizations of a collection of iid RVs (the same as realizing the same RV many times).

*Technically random sample is only iid RV if replacement, however if population is large enough, sample without replacement closely approximates with (as shown in discussion 3).

Random Variables in Sampling and Estimation

Motivating Example:

A car manufacturer uses an automatic device to apply paint to engine blocks. Since engine blocks get very hot, the paint must be heat-resistant, and it is important that the amount applied is of a minimum thickness. A warehouse contains thousands of blocks painted by the automatic device. The manufacturer **wants to know the average amount of paint applied by the device**, so 16 blocks are selected at random, and the paint thickness is measured in mm. The results are below:

1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83, 1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49

We can estimate

μ: the population **mean** paint thickness of all blocks using the
sample mean $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ as an _estimator_ of the **true mean** (μ)
The calculated **statistic** _(RV)_ from the sample data is $\hat{\mu} = \bar{X} = 1.348$

$\hat{\mu}$ = estimate
$\bar{x}$ = realization

We'd expect this estimate to be _different_ in another sample. We need to understand the _variability_ of the $\hat{\mu} = \bar{X}$ in repeated sampling so we know how precisely we can estimate μ from our estimator $\hat{\mu}$**.**

Random Variables in Sampling and Estimation

The sample mean is a linear combination of 16 iid random variables. Each measured value is a realization of the RV and once the statistic is computed from values it is called an __estimate__.

There are many ways to define an estimator $\hat{\theta}$ of any given population parameter θ. Because estimators are random variables, we can consider their distribution's __shape__, expectation and variance to determining which is __best__ for our purposes.

The probability distribution of a statistic $\hat{\theta}$ is called its **sampling distribution.**

*it shows the variety of values the statistic can take over __repeated__ sampling

*$E(\hat{\theta})$ is the mean of all possible values of $\hat{\theta}$.

mean of statistic

true mean

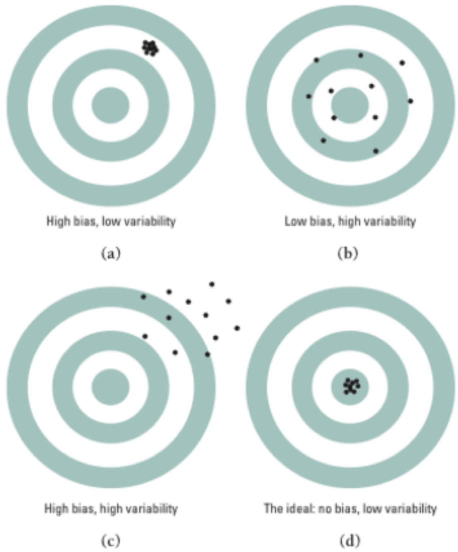
In **unbiased** estimators, $bias(\hat{\theta}) = E(\hat{\theta}) - \theta = 0$

* Var$(\hat{\theta})$ is the variance of all possible values of $\hat{\theta}$

$\sqrt{\mathrm{Var}(\hat{\theta})}$ is called the __Standard error__ $\mathrm{SE}(\hat{\theta})$

* mean squared error: $\mathrm{MSE}(\hat{\theta}) = Var(\hat{\theta}) + bias(\hat{\theta})^2$

Bias, variability, and shape We can think of the true value of the population parameter as the bull's- eye on a target and of the sample statistic (estimate) as an arrow fired at the target. Both bias and variability describe what happens when we take ___many___ shots at the target.

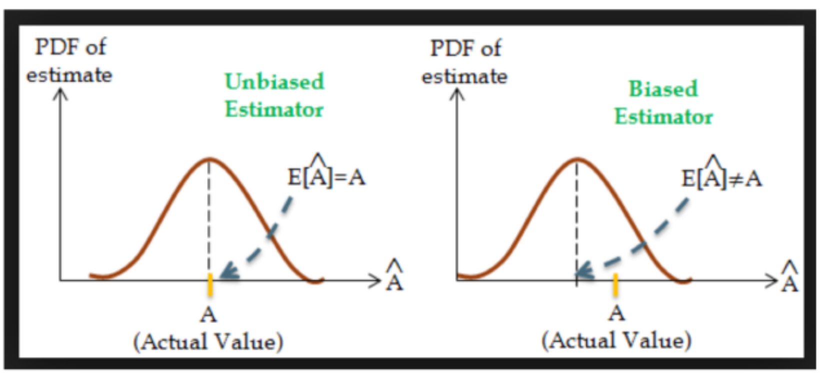


Bias means that our aim is off and we consistently miss the bull's-eye in the ___direction___. Our sample values do not center on the population value.

Variability means how widely scattered the different statistic values are. High variability means repeated samples do not give very ___similar___ results.

The lesson about center and spread is clear: **given a choice of estimators to estimate an unknown parameter, choose one with ___no or low___ bias and minimum ___variability___ (there is often a tradeoff)**.

Bias, variability, and shape of an estimator's distribution



Bias means our sample values do not ___center on the population value___

Repeated Sampling...

Variability High variability means repeated samples do not give very similar results.

Sampling Distribution of the Sample Mean (for a known population)

X	P(X=x)
0	0.2
3	0.3
12	0.5

A large population is described by the probability distribution

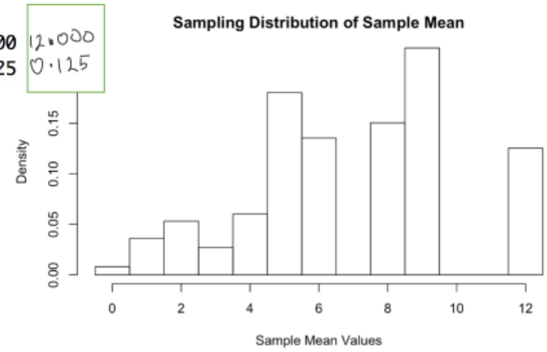
with, $E(X) = \mu = 6.9, \quad Var(X) = \sigma^2 = 27.09$

Let X1, X2, X3 be a random sample of size 3 from the population

(because of "large" population, we will compute as if independent).

a. Determine the sampling distribution of the sample mean $\bar{X}$ by listing all possible samples and their probabilities. (see R code)

```
unique_means 0.000 1.000 2.000 3.000 4.00 5.00 6.000 8.00 9.000 12.000
tot.prob     0.008 0.036 0.054 0.027 0.06 0.18 0.135 0.15 0.225 0.125
```



Sampling Distribution of the Sample Mean (for a known population)

X	P(X=x)
0	0.2
3	0.3
12	0.5

A large population is described by the probability distribution:

with, $E(X) = \mu = 6.9, \quad Var(X) = \sigma^2 = 27.09$

Let X1, X2, X3 be a random sample of size 3 from the population

a. Calculate $E(\bar{X})$, Var($\bar{X}$) and SE($\bar{X}$) using the sampling distribution, and then again using our random variable rules.

```
X̄          unique_means 0.000 1.000 2.000 3.000 4.00 5.00 6.000 8.00 9.000 12.000
P(X̄=x̄)    tot.prob     0.008 0.036 0.054 0.027 0.06 0.18 0.135 0.15 0.225  0.125
```

$E(\bar{X}) = E\left(\frac{X_1+X_2+X_3}{3}\right) = \frac{1}{3}\left[E(X_1)+E(X_2)+E(X_3)\right] = \frac{1}{3}\left[6.9+6.9+6.9\right] = 6.9$

$\text{Var}(\bar{X}) = var\left(\frac{X_1+X_2+X_3}{3}\right) = \frac{1}{9}\left[var(X_1)+var(X_2)+var(X_3)\right] = \frac{1}{9}\cdot 3\cdot\sigma^2 = \frac{\sigma^2}{3} = \frac{27.09}{3}$

$E(\bar{X}) = \sqrt{\frac{\sigma^2}{3}} = \frac{\sigma}{\sqrt{3}}$

$Var(\bar{X}) =$

Combinations of Random Variables continued

In general, suppose we have

iid random variables: **X1, X2, X3, ... Xn,** with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

Consider the distribution of the Mean of those variables $\bar{X} = \frac{X_1+X_2+..X_n}{n}$

$E(\bar{X}) = E\left(\frac{X_1 + X_2 + ..X_n}{n}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) ... + E(X_n)\right] = \frac{1}{n} n E(X_1) = E(X_1) = \mu$

$Var(\bar{X}) = Var(\frac{X_1+X_2+..X_n}{n}) = \frac{1}{n^2} Var(X_1 + X_2 + X_3 ... + X_n) = \frac{1}{n^2} n \, Var(X_1) = \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$

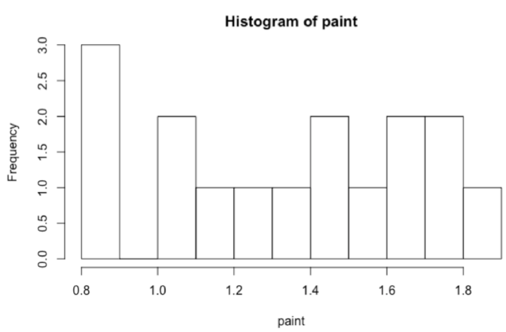
Notice, the sample mean is unbiased and the variability of the sample mean decreases as our sample size increases ...(why?), What also affects variability of the mean?

The mean squared error $MSE(\bar{X}) = Var(\bar{X})$+bias$(\bar{X})^2 = Var(\bar{X})$

If **X1, X2, X3, ... Xn** are Normal, we know $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Assessing Normality Assumption

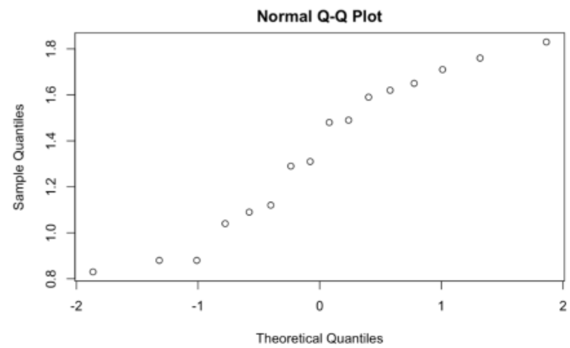
- How do we know if our 16 paint observations (or any others?) are RV from a Normal Population? – we can't ever "know" for sure,
  - how Normal is "Normal-enough"?
- A histogram of the sample data **may** show a similar 68-95-99.7 pattern, but with small samples and sample variability, a bell-shaped curve isn't always apparent, even when sampling from a Normal distribution.



Assessing Normality Assumption

- How do we know if our 16 paint observations (or any others?) are RV from a Normal Population? – we can't ever "know" for sure
  - how Normal is Normal-enough?
- A **normal quantile-quantile plot** or __normal QQ plot__ depicts the $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$ quantiles from $N(0, 1^2)$ on the x axis

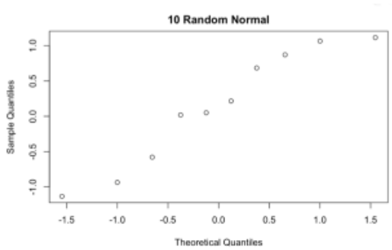
against the sorted data set on the y . The ideas is that, the more the points resemble a __straight line__, the stronger evidence we have that they came from the same distribution (ie that our sample Came from a normal distribution).

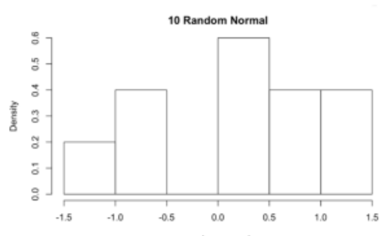


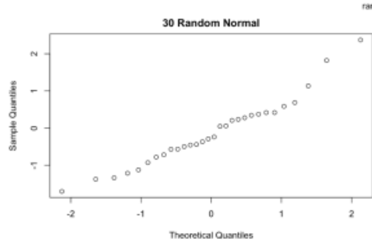
Assessing Normality Assumption – QQ Plots

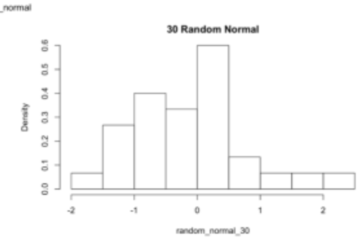
But there is large variability in qqplot we see (even if we are sampling from a __normal__ distribution)– especially with __small__ samples – try for yourself in R!

Sampling from a Normal(0,1) population:





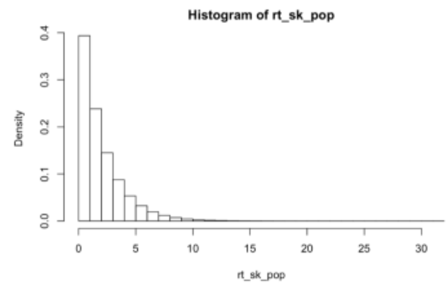


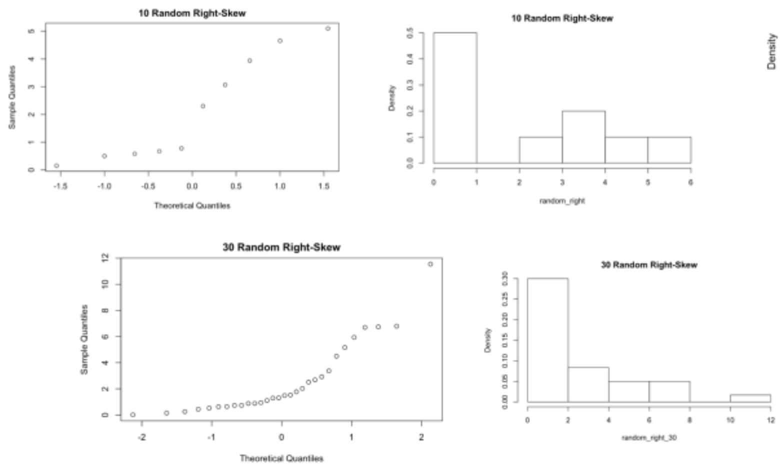


Assessing Normality Assumption – QQ Plots

But there is large variability in qqplot we see (even if we are sampling from a normal distribution)– especially with small samples – try for yourself in R!

Sampling from a right-skewed population:

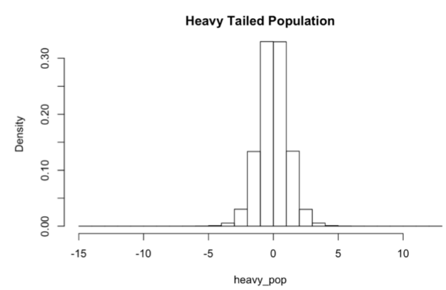


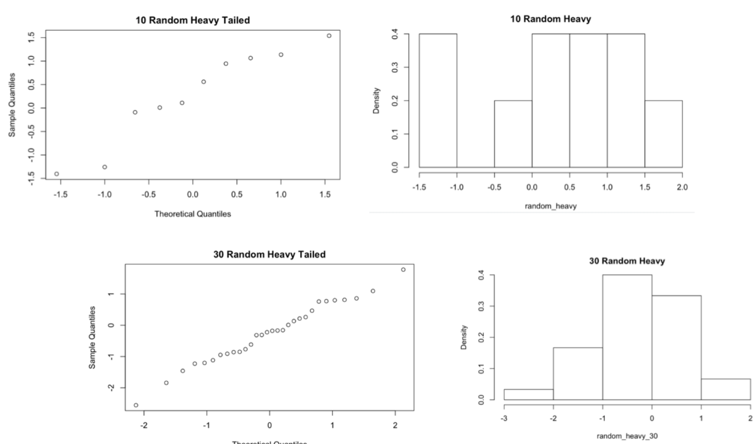


Assessing Normality Assumption – QQ Plots

But there is large variability in qqplot we see (even if we are sampling from a normal distribution)– especially with small samples – try for yourself in R!

Sampling from a “heavy-tailed” population:

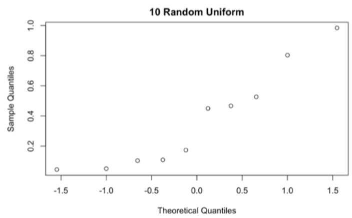


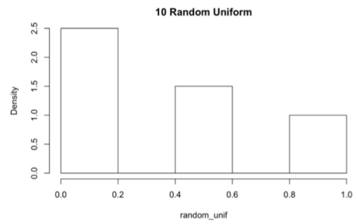


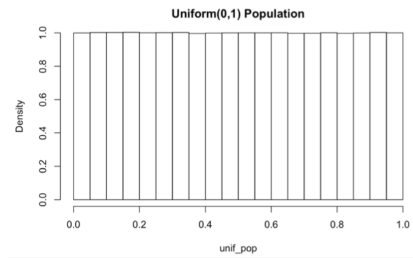
Assessing Normality Assumption – QQ Plots

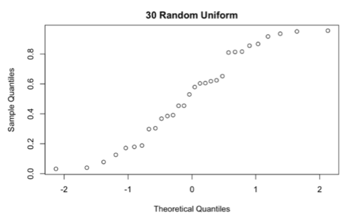
But there is large variability in qqplot we see (even if we are sampling from a normal distribution)– especially with small samples – try for yourself in R!

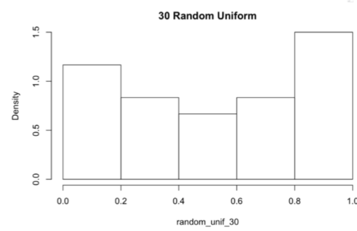
Sampling from a Uniform(0,1) population:











The Central Limit Theorem

"Whatever the population with finite mean and variance, the distribution of $\bar{X}$ is approximately Normal when __n is large enough__"

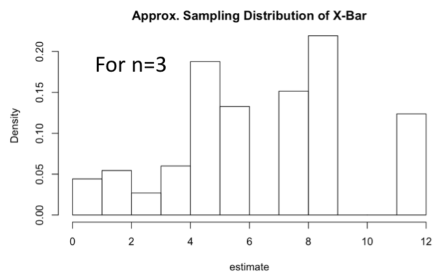
Let X1, X2, …Xn be a collection of iid RVs with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. For **large enough n**, the distribution of $\bar{X}$ will be approximately normal with $E(\bar{X}) =$ __μ__ and $\mathrm{Var}(\bar{X}) =$ __$\frac{\sigma^2}{n}$__.

The more "Non Normal" the population is, the __larger__ the sample size n needs to be for the mean to be approximately $\mathrm{N}(\mu, \frac{\sigma^2}{n})$.

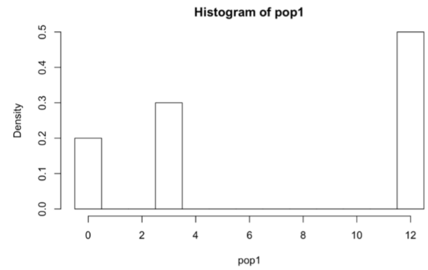
(n >30 often counts as "large enough" for roughly symmetric distributions)

Central limit theorem seen in Non-normal populations

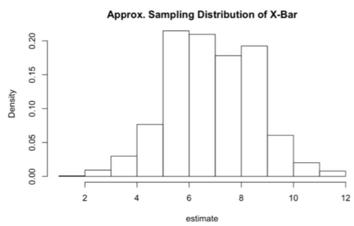
A large population is described by the probability distribution with, $E(X) = \mu = 6.9, \quad Var(X) = \sigma^2 = 27.09$



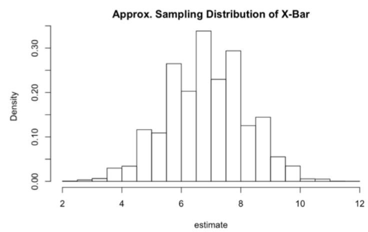
X	P(X=x)
0	0.2
3	0.3
12	0.5



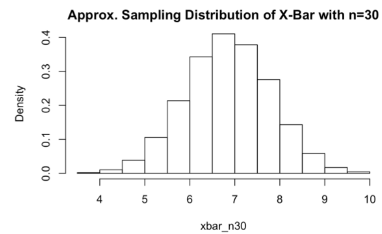
For n=10



For n=15



For n=30



So we have two situations in which $\bar{X} \sim N\left(\mu, var = \frac{\sigma^2}{n}\right)$

1. The population is Normal: $N(\mu, \sigma^2)$ (then " ~ " is exact for any size n)
2. The sample size n is large enough that CLT applies (" ~ " is approximate)

- Ex 1: Suppose we are taking a random sample of 45 elements from the last population X is non normal; $E(X) = \mu = 6.9, \quad Var(X) = \sigma^2 = 27.09$.
  - Describe the shape, expectation, and variance of the sampling distribution for the mean estimator
  - Calculate the probability that the mean of the sample we observe will be greater than 7.

symmetric and not skewed

$\bar{X} \approx$ Normal because CLT (similar size of 30 big enough)

want to the right so 1−

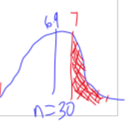
$E(\bar{X}) = 6.9$

$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{27.09}{45} = .602$

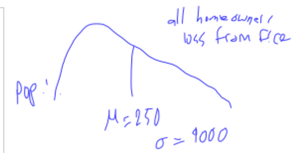
$P(\bar{X} > 7) = P\left(z > \frac{7-6.9}{\sqrt{.602}}\right) = P\left(z > \frac{7-6.9}{.776}\right)$

$= P(z > 0.129) = 1 - .5517$

$= \boxed{.4487}$



- Ex 2: An insurance company knows that in the population of millions of homeowners, the mean annual loss from fire is μ = \$250 and the standard deviation is σ = \$1000. (The loss distribution is strongly right-skewed, since most policies have no loss but a few have large losses.) If the company sells 10,000 policies, can it safely base its rates on the assumption that the average loss will be no greater than \$275? What assumptions does the company have to make?



$P(\bar{X} < 275)$ or $P(\bar{X} \le 275)$ because continuous

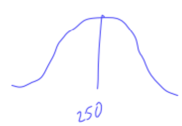
$P\left(z < \frac{275 - 250}{10}\right) = P(z < 2.5) = .994$

$\bar{X} \sim N(250, 10)$

n is large
10,000
CLT can overcome skew of pop

$E(\bar{X}) = 250$

$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{\sigma}{\sqrt{n}} = \frac{1000}{\sqrt{10{,}000}} = 10$



For Next Time

- Start working through posted homework. Post questions on Piazza.