Teaching Introductory Statistics with R/RStudio

Probability Distributions

Central Limit Theorem

Probability Distributions in R

Four Functions that can be used for any Distribution:

`p` → Probability/Proportion/Area Under Distribution Curve (Integral)

`q` → Quantile (Decimal form of Percentile)

`d` → Density (y-coordinates of points on the Density Curve)

`r` → Random Draw from the Distribution

Put one of these letters in front of the density name.
Example - Normal Distribution: `pnorm()`, `qnorm()`, `dnorm()`, `rnorm()`

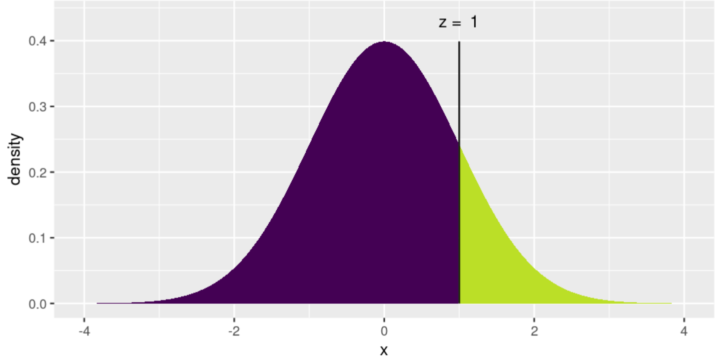
Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

https://www.stat.umn.edu/geyer/old/5101/rlook.html

xpnorm()

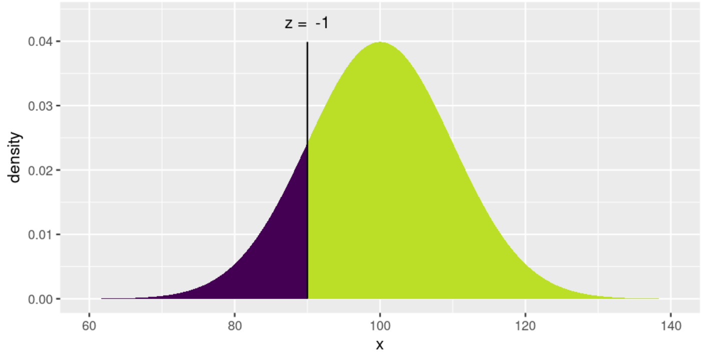
Z-Scores: Values in the standard normal distribution, mean is 0, standard deviation is 1. Find the probability of obtaining a z-score less than 1.

```
xpnorm(1, mean = 0, sd = 1, lower.tail = TRUE)
[1] 0.8413447
```



A population has a mean of 100 and standard deviation of 10. Find the probability that a randomly selected member of this population has a value of 90 or greater.

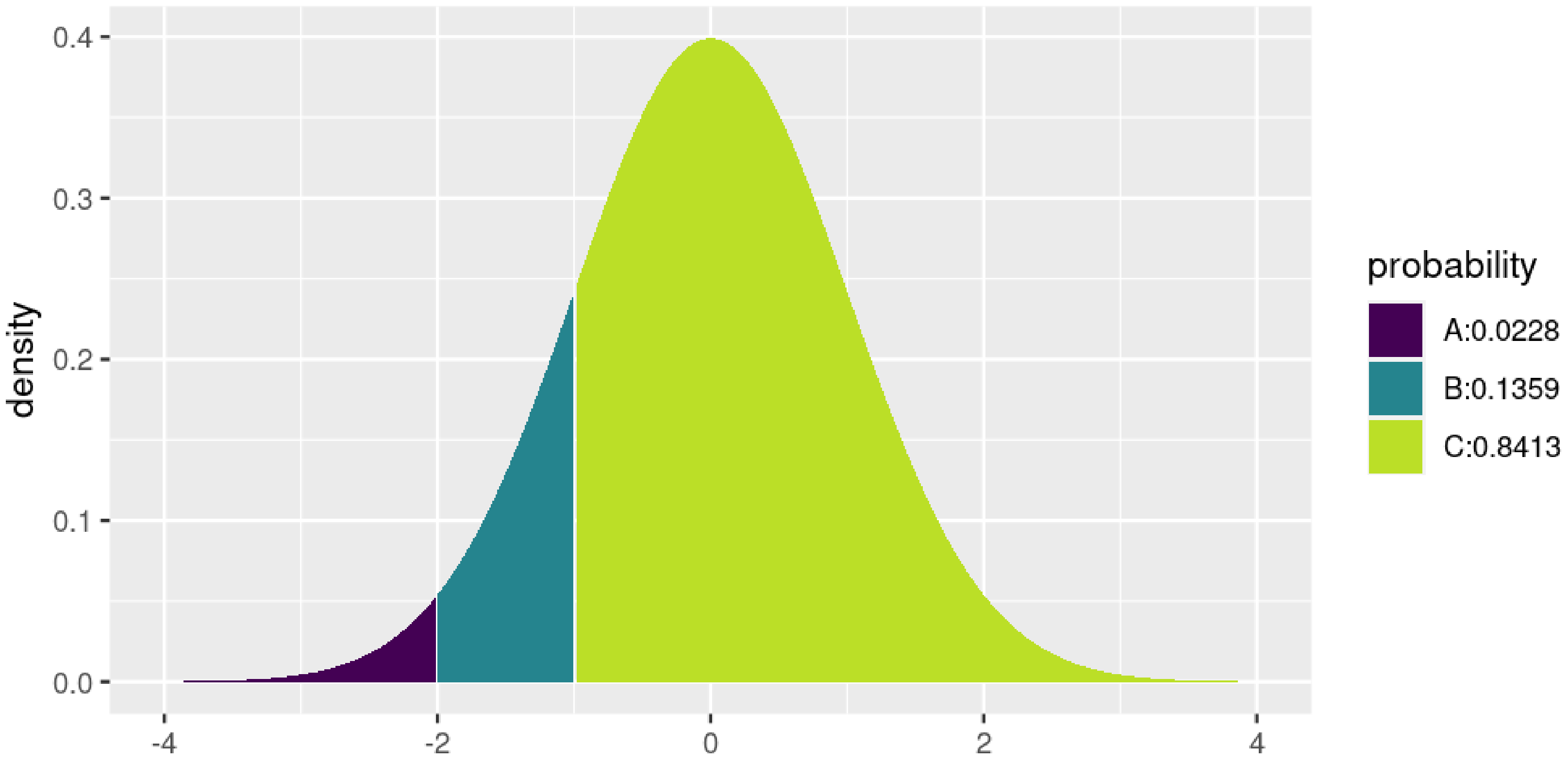
```
xpnorm(90, mean = 100, sd = 10, lower.tail = FALSE)
[1] 0.8413447
```



xpnorm()

Find the probability of obtaining a z-score between -2 and -1.

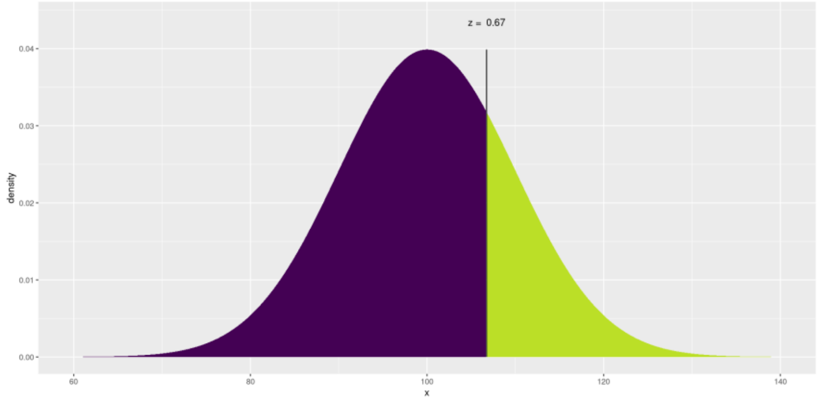
```
xpnorm(c(-2,-1), mean = 0, sd = 1)
[1] 0.02275013 0.15865525
```



xqnorm()

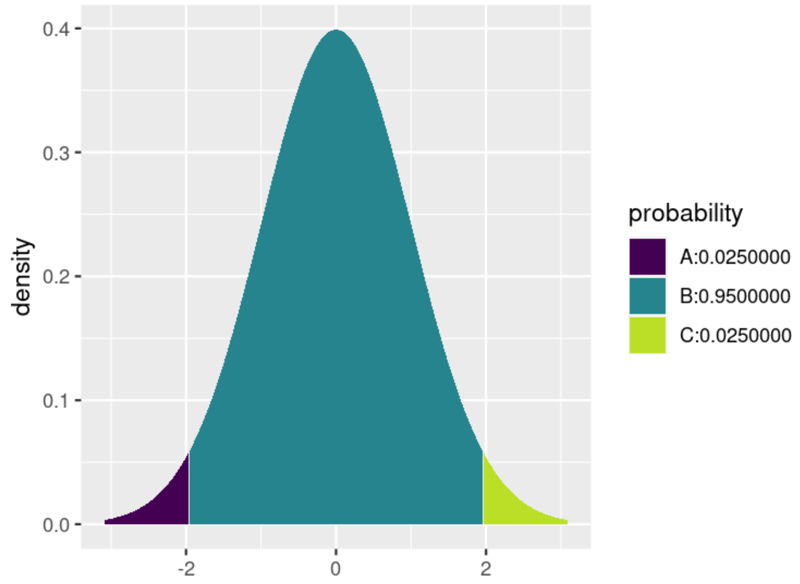
A population has a mean of 100 and standard deviation of 10. Find the 75th percentile.

```
xqnorm(0.75, mean = 110, sd = 10, lower.tail = TRUE)
[1] 0.6744898
```



Which two z-scores cut out the middle 95% of the standard normal distribution?

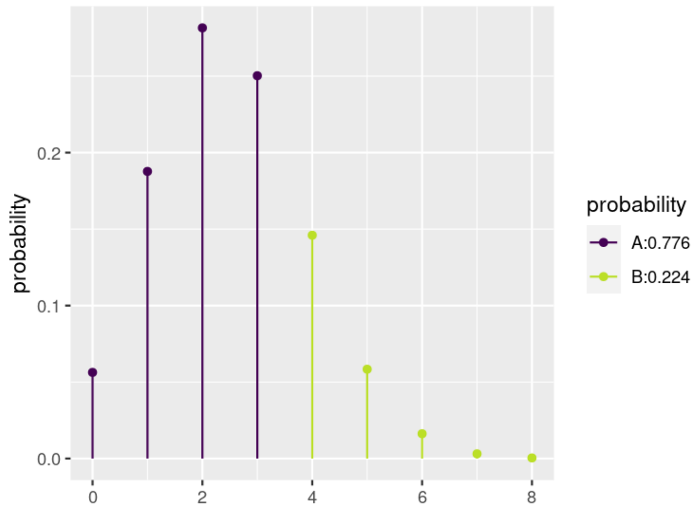
```
xqnorm(c(0.025,0.975), mean = 0, sd = 1)
```



Other examples – Binomial and t

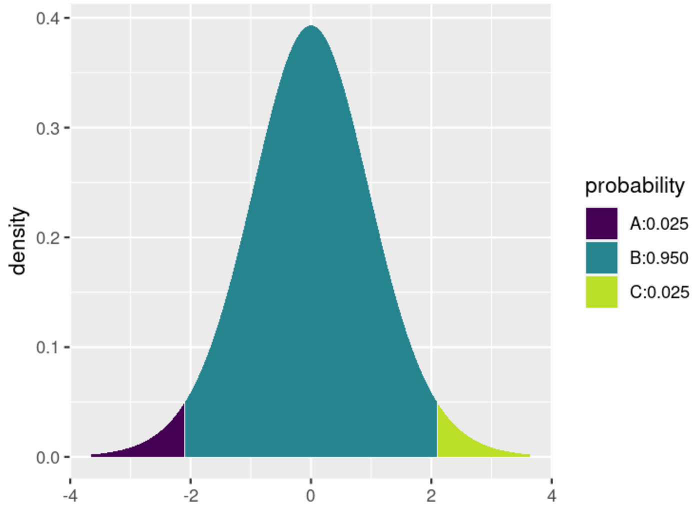
Binomial Experiment: Find the probability of at most 3 successes in 10 trials given a success probability p = 0.25.

```
xpbinom(3, size = 10, p = 0.25, lower.tail = TRUE)
[1] 0.7758751
```



Which two t-scores cut out the middle 95% of the t-distribution with 17 degrees of freedom?

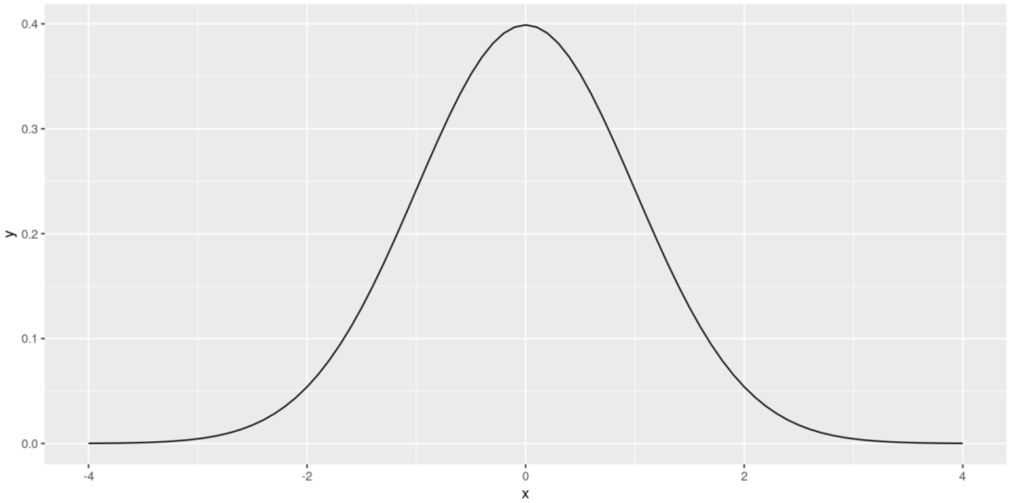
```
xqt(c(0.025,0.975),df = 17)
[1] -2.109816  2.109816
```



d (Density) and r (Random)

Plot The Z-Curve. Generate points, connect with line.

```
> x <-seq(from = -4, to = 4, by = 0.1)
> y <-dnorm(x, mean = 0, sd = 1)
> gf_line(y ~ x)
```



One Random draw from a Uniform distribution Unif(0,1)

```
> runif(1, min = 0, max = 1)
[1] 0.6123176
```

Random sample of size 5 from a Normal distribution N(0,1)

```
> rnorm(5, mean = 0, sd = 1)
[1]  0.5920287  0.4656121 -2.0115025  0.3550288  1.1795863
```

Simulation in R – Central Limit Theorem

SIMULATION EXPERIMENT: Understanding the variability of a point estimate

Suppose we knew that the true proportion of **ALL** American adults who support expansion of solar energy is known to be 88%, so $p = 0.88$ is the population parameter.

If we poll 1000 U.S. adults, how close would sample proportion $\hat{p}$ be to the true parameter p = 0.88?

If we take many random sample, what would the distribution of all the $\hat{p}$'s look like?

CLT Simulation in R: Understanding Variability of a Point Estimate

Poll 1000 Adults, calculate proportion of support. Repeat 500 times, get 500 sample proportions!

```
# Let "1" represent "support", "0" represent "not support"
pop_size <-250000000
# Let's make a vector of the population opinions with 88% 1's and 12% 0's
opinions <-c(rep(1, 0.88*pop_size), rep(0, 0.12*pop_size))
# Let's take a sample of size 1000 from the population of opinions
s <-sample(opinions, size = 1000)
# Let's find the proportions for our sample
prop(s)
# Let's repeat 500 times the process of sampling 1000 opinions.
# We'll store in the results data frame, where each row is a sample of 1000
# dimensions of data frame is 500 rows, 1000 columns:
df_sample <- do(500) * sample(opinions, size = 1000)
dim(df_sample)
df_sample
# Let's now find the proportion for each sample. We do this by applying the prop function to each row.
# Then we'll store the proportions from all 500 samples in a vector.
prop_support <-apply(df_sample, prop, MARGIN = 1)    #1 for rows, 2 for columns.
# See distribution of these proportions. Use gf_dhistogram for density. Then overlay with normal distribution
gf_dhistogram(~ prop_support, bins = 12, color= "black", fill = "gold") %>%
       gf_fitdistr(dist = "dnorm")
```

CLT Simulation in R: Understanding Variability of a Point Estimate