Stats Overview

Contents

Libraries Used

```
library(gmodels)
library(PSYC201)
library(ggplot2)
library(dplyr)
library(magrittr)
library(moments)
library(combinat)
library(gtools)
```

Permutation Rules

Number of permutations with replacement, any possible order

$$n^r$$

: n is number of items to choose from, r is number chosen- makes sense.

Total Number of permutations without replacement

$$\frac{n!}{(n-r)!}$$

-basically a function that gives you limit on declining number of possible options

Number of permuatations without replacement, but not requiring a specific order

$$\frac{n!}{r!(n-r)!}$$

- it is the above, but now we need to take into account the fact that the $r $ selected can come in any order.

Univariate descriptive Statistics

Definition of moments.

Point of clarification: Moments refers just the the set of numbers that describe a set of points, it makes sense in a physical system or to probability. They are defined according to the following function.

$$\mu_i' = \int_a^b (x-c)^i f(x) dx$$

For probability $f(x) = p(x)$ so :

$$\mu_i' = \int_a^b (x-c)^i p(x) dx$$

As i increases, you move up moments. The total number of moments as i goes to inf completely describes a set of points. *Seems like you need might only need n moments, but what do I know.* In general, each moment is referred to as μ_i, my guess is that using μ to represent mean is a convention to drop the 1.

Zeroeth Moment- the probability mass function

$M_0 = \mu_0 = \Omega = 1$

First moment, the mean

$M_1 = \mu_1 = \mu_x = \int_a^b (x-c)^1 f(x) dx = \int_a^b x f(x) dx = E[X]$

Second Moment, variance. Note that σ is the standard deviation.

$M_2 = \mu_2 = \sigma_x^2 = \int_a^b (x-c)^2 f(x) dx = E[(X - E[X])^2] = E[X^2] - E[X]^2$

Confirming with R

```r
# need to adjust for sample calculation by subtracting 0, but not identitcal
x <- rnorm(200, 0, 4)
var(x)
```

```
## [1] 16.54169
```

```r
sum((x^2))/(length(x) - 1) - mean(x)^2
```

```
## [1] 16.54298
```

Why don't I get the same results

Need to include a hardcoded function for evaluating variance of discrete function, maybe break it out for kurtosis and skew calculations

This continues for the definition of moment, but using these in an non-normalized fashion doesn't appear normal

Third moment

$M_3 = \mu_3 = \int_a^b (x-c)^3 f(x) dx$

Forth Moment

$M_4 = \mu_4 = \int_a^b (x-c)^4 f(x) dx$

At this point, there now are what called "normal moments", where the are now "normliazed" by dividing by the standard deviation. Now the i_{th} moment is given by

$$\frac{\mu_i}{\sigma^i}$$

The third normalized or standardized moment is called Pearson moment coefficient of skeweness.

$M_3 = \gamma = \frac{\mu_3}{\sigma^3} = E[\left(\frac{s - \mu_1}{\sigma}\right)^3] = E[\left(\frac{(E[(X-\mu)^3]}{E[(X-\mu)^2]^{\frac{3}{2}}}\right)$

Confirming with r

```r
x <- rgamma(200, shape = 10)
skewness(x)
```

```
## [1] 0.5260652
```

```r
mean(mean((x - mean(x))^3)/(mean((x - mean(x))^2)^(3/2)))
```

```
## [1] 0.5260652
```

Forth Moment, kurtosis.

$M_4 = \beta_2 = \frac{\mu_4}{\sigma^4} = E[\left(\frac{(E[(X-\mu)^4]}{E[(X-\mu)^2]^2}\right)]$

Confirming with r

```r
x <- rnorm(200)
kurtosis(x)
```

```
## [1] 2.761481
```

```r
mean(mean((x - mean(x))^4)/(mean((x - mean(x))^2)^(2)))
```

```
## [1] 2.761481
```

However, because the kurtosis is equal to three in a normal distribution, is is often subtracted from the measure of kurtosis so that the sign of the value indictes whether a distribution is platykurtic (less than 0, low kurtosis) or leptokurtic (greater than 0, high kurtosis). It is simultaneously a measure of three values. It generally indcreases with peakedness, how much variance is right next to the mean and tails and the tail thickness, but decreases with increased mass held in the sholders, or more modest deviations. However, it generally is seen as mostly a measure of tail thickness.

$$M_4 = \beta_2 = \frac{\mu_3}{\sigma^3} = E[\left(\frac{(E[(X-\mu)^4]}{E[(X-mu)^2]^2} \right) - 3]$$

Central Limit Theorem.

Proof by assertion. It's an interesting idea. What happens when you start to sum n, iid variables. This is interesting as it assumes it observation is itself drawn from an independent distribution (which i guess is restating it's assumption). If you know the nature of the underlying distribution, you can calculate your expectation about the mean, variance, skew and kurtosis of the sum of the variables.

Rules for adding or summing variables.

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

Question:

Var(X+Y) = Var(X) + Var (Y)
if Y = X
why does Var(2*X) =! Var(X + X)? Because they are different transformations, doing something very different. This is a dumb question rob.

This means that if one starts to sum iid varables, n increases linearly. However, by assertion the kurtosis and skew start to fall towards zero, meaning that distribution of the sum of iid variables starts to look like a normal distribution. Apprently they increase like this, regardless of the initial distribution.

$$Mean[\sum^{n} X] = n * Mean[X]$$

$$Variance[\sum^{n} X] = n * Variance[X]$$

$$skew(\sum^{n} X) = \frac{1}{\sqrt{n}} * skew(x)$$

$$kurtosis(\sum^{n} X) = \frac{1}{n} * kurtosis(x)$$

This more or less underlies most statistics. There is the assumption that as you increase n, the expectation about the sum of the added random variables starts to resemble a normal variable (though normally you divide this value by n). This is cool.

The actual Central limit theorem

$P(\bar{(}X)_n) \rightarrow N(\mu_X, \frac{\sigma_X}{\sqrt{n}}$

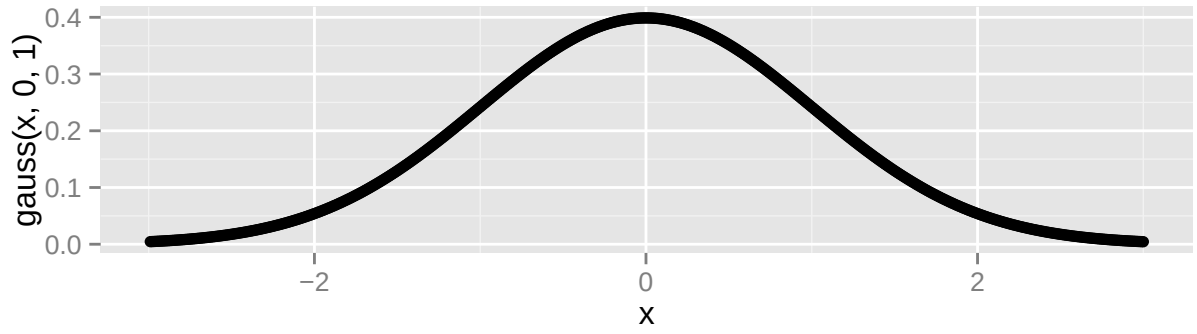
Normal variable

Here X is distributed according to the following function.

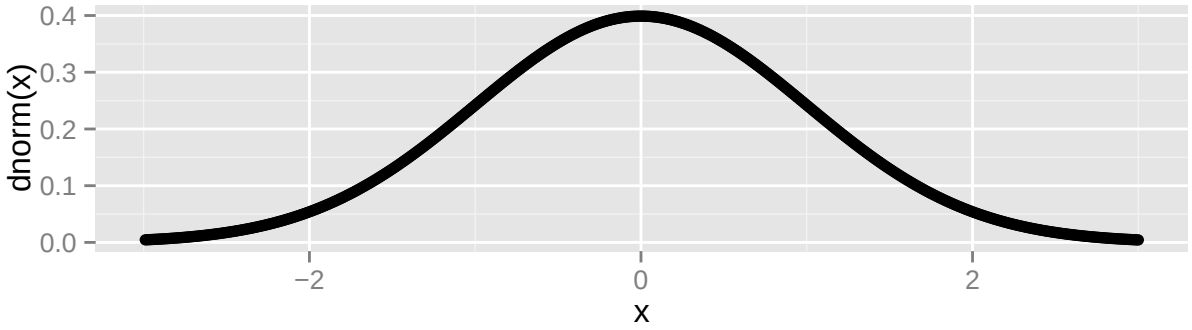
$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e\left[-\frac{x-\mu}{2\sigma^2}\right]$$

Check in R

```
gauss <- function(x, mu, sigma) {
    const <- 1/sqrt(2 * pi * sigma^2)
    probability <- const * exp(-(x - mu)^2/(2 * sigma^2))
    return(probability)
}

x <- 1:600/100 - 3

qplot(x, gauss(x, 0, 1))
```




```
qplot(x, dnorm(x))
```



Laws of large numbers

If you have now taken the sum of this set of random variables, you can now get the expected value of the mean and variance by diving by n. I think this means that there is a careful set

- Strong Law of large numbers

$$\bar{X} \to Mean[X]$$

- Weak Law of large numbers

$$Prob(\bar{X} - \mu > e) \to 1$$

- Borels Law of large numbers

$$\frac{\sum i = 1^n x}{n} \to P(X = x)$$

- Monte Carlo principle

$$\frac{1}{n} \sum i = 1^n f(x^i) \to E_{P(x)}[f(X)]$$

Multivariate extension

Fill in later

Sampling Theory

Expected Value of the mean

Knowing the following

$$Mean[\sum^n X] = n * Mean[X]$$

$$Variance[\sum^n X] = n * Variance[X]$$

$$skew(\sum^n X) = \frac{1}{\sqrt{n}} * skew(x)$$

$$kurtosis(\sum^n X) = \frac{1}{n} * kurtosis(x)$$

What is the distribution we can expect when we sum X iid variables and divide by n. This is hte sampling distribution of the sample mean. Sample mean of n iid variables is is viewed as the distribution that results from the following. $\frac{X_1 + X_2 + ... + X_n}{n}$

So sum up the distributions, divide the sums by n, which gives us the sampling distribution

$$Mean[\frac{1}{n} * \sum^n X] = Mean[X]$$

$$Variance[\frac{1}{n} * \sum^n X] = \frac{1}{n^2} * n * Variance[X] = \frac{1}{n} * Variance[X]$$

If there is time, solve the following. Right now, just take as an article of faith that they converge rapidly to the parametrs of a normal distribution.

$$skew(\frac{1}{n} * \sum^n X) =?$$

$$kurtosis(\frac{1}{n} * \sum^n X) =?$$

What this means is the central limit theorem.

$$P(\bar{(}X)_n) \to N(\mu_X, \frac{\sigma_X}{\sqrt{n}}$$

Z (Standard Normal) distribution

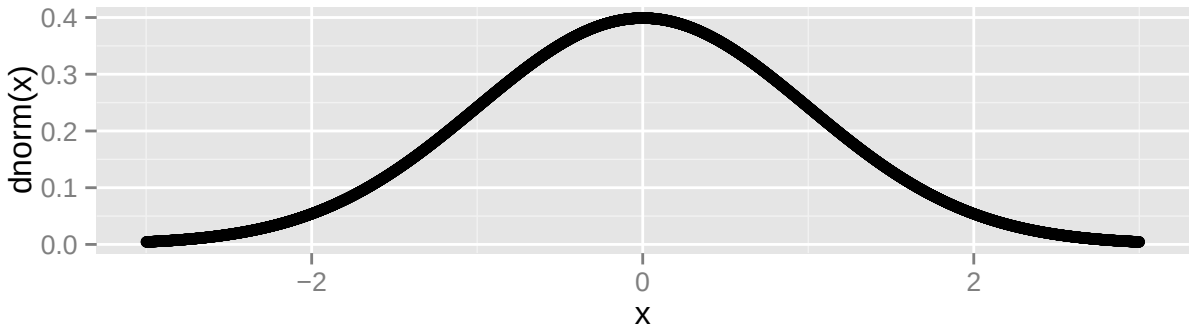
The $N(0,1) distribution. Can transform distribution x to a z distribution with the following forumula $z_x = (x - \mu_x)/sd_x$ Subtract the population mean, divide by the population standard deviation.

Particularly valuable because it allows one to compare different distributions. Often one looks at the z score transform of the sample mean, as it lets one run tests of it's value.

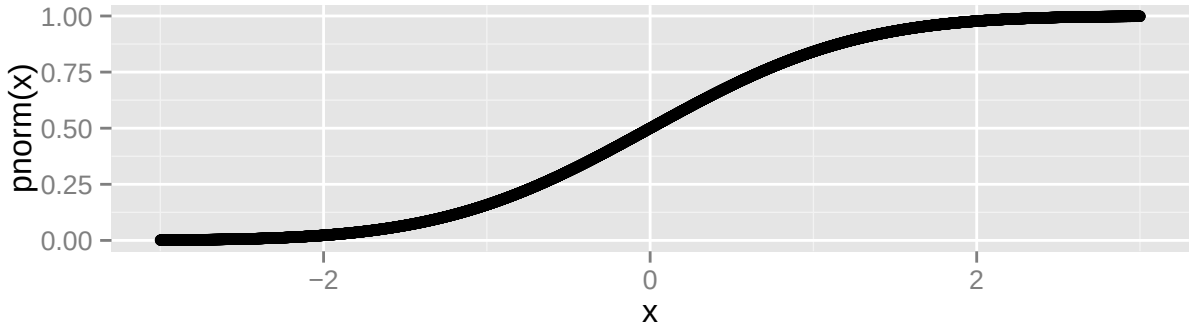
pdf,cdf, icdf

Here are the r commands, same for continuous and discrete probability density functions.

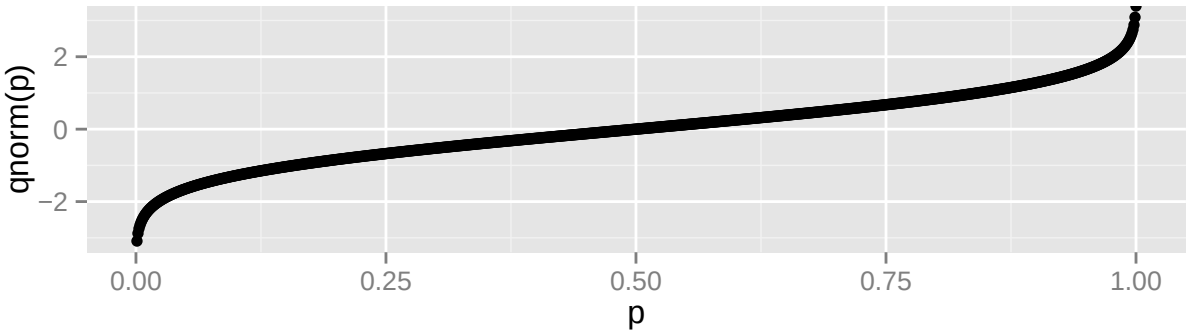
```
# pdf Probability mass function - what is the probability of this specific
# point, not really interpretable in a continuous distribution, but useful
# for discrete distributions.
x = 1:6000/1000 - 3
qplot(x, dnorm(x))
```




```
# cdf Cumulative Density Function what is the probability of observing this
# this value or any value that is lower. This is inclusive the x value,
# which is important to remember for discrete distributions
x = 1:6000/1000 - 3
qplot(x, pnorm(x))
```




```
# icdf,Inverse Cumulative Density function or quatile distribution. With
# this distribution, it returns the value such that p% of the probability
# distribution is lower than that value.
p <- 1:1000/1000
qplot(p, qnorm(p))
```



-Not refer to law of total probability in the answers, particularly as numerator in bayes equation.

Null Hypothesis Sampling Testing.

Confidence intervals

- Identify the regions where there is an X% of the probability falls inside the value of that confidnece interval.
- Are weird and non-intuitive
  - sample mean is random variable
  - population mean is fixed point value
  - can't say anything about probility distribution of population mean
  - can say- if we drew infinite sample of this size and calculated the 95%CI, then 95% of these intervals will cover sample mean
  - This is referred to as coverage, and frequentist statistics is optimized for this value.
- Frequentist procedures more concerned with procedure
- The Bayesian credible interval, under reasonable assumptions very closely resembles the confidence interval.

Quick sanity check

```
x <- rnorm(10000,3,5)
#95CI on the sample error
meanx <- 3
varx <- 5^2/10000
sdx <- 5/sqrt(10000)
#lower bound
qnorm(.05/2,meanx,sdx)
```

```
## [1] 2.902002
```

```
#upper bound
meanx+(meanx-qnorm(.05/2,meanx,sdx))
```

```
## [1] 3.097998
```

```r
ci(x)
```

```
##   Estimate   CI lower   CI upper Std. Error
## 2.93739033 2.83873362 3.03604705 0.05032989
```

Significance testing.

Terms

Alpha - α - probability that you will observe outcomes more extreme than one's cuttoff for significance, conditional on the null hypothesis being true. Estimated frequency of Type 1 error.

Cohen's d $= d = |\frac{\mu_T - \mu_0}{\sigma_X}|$ - measure effective size generally use $\bar{x}$ for μ_t.

Power, or Beta - β Probability of accepting rejecting null hypothesis, conditional on the alternative or tested hypothesis being true. Consequence of frequentist statistics. One identifies a value with alpha, where an observed value closer to the mean of the null hypothesis means the null hypothesis is accepted.

Strategy for Calculating Power

```r
# null hypothesis parameters
mu0 <- 5
sd0 <- 5
# Alternative hypothesis parameters
muT <- 10

n <- 25
alpha = 0.05

# set up limits for null hpyothesis
dif <- abs(mu0 - qnorm(alpha/2, mu0, sd0/sqrt(n)))
uplim <- mu0 + dif
lowlim <- mu0 - dif

# now look at probability of seeing values outside that confidence interval,
# this is power

above <- 1 - pnorm(uplim, muT, sd0/sqrt(n))
below <- pnorm(lowlim, muT, sd0/sqrt(n))

EstPower <- above + below
```

Tricks for estimating power

```r
# check given above
d <- abs((mu0 - muT)/sd0)


# power given effect size, it works but it is ignoring the probabiltiy that
# it will be rejected in the wrong way
pow = 1 - pnorm(abs(qnorm(alpha/2)) - d * sqrt(n))
pow == above
```

```
## [1] TRUE
```

```r
# number needed for n- this doesn't include 'other sideded power, but maybe
# not important'
n.needed = (abs(qnorm(alpha/2) - qnorm(pow))/d)^2
```

T Test

Necessity

Basically using a systematically biased estimate of standard deviation leads to really wrong estimates of signifcance and is hard to correct.

By Assertion, we start to correct this by estimating the population variance as

$$\hat{sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Now the population standard devation is

$$\hat{sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x})^2}$$

Then you could use the sample standard deviation (note, not the sample standard error) to calcualte z score

$$Z_x = \sqrt{n}(\frac{\bar{x} - \mu_0}{s})$$

note this framing makes sense- you multiply by the square root of n, becuse the z score for the difference between the observed mean and the null mean increases with sample size.

```r
x <- rnorm(10000, 5, 10)
sqrt(10000) * (5 - 0)/10
```

```
## [1] 50
```

```r
(5 - 0)/(10/sqrt(10000))
```

```
## [1] 50
```

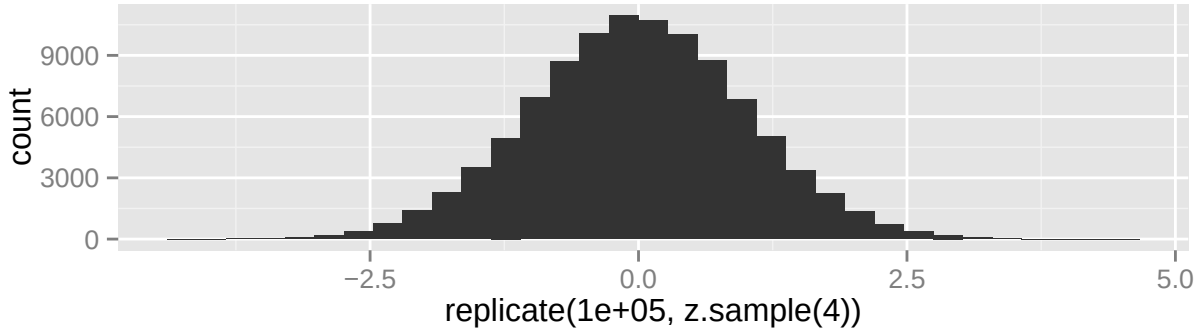
However, this doesn't work b/c the the z score calculated this way doesn't follow normal distribution, because the variance has sampling variation. This means statistics calculated from the z. What the above calculation actually gives you is the t score

$$t_{\bar{x}} = \sqrt{n}(\frac{\bar{x} - \mu_0}{s_x}) == dfrac\bar{x} - \mu_0 s_x/\sqrt{n}$$

You can see the problem if you hardcode the standard deviation, here at 1 as is dones in the z sample, vs estimate the standard devaition with the sd function.

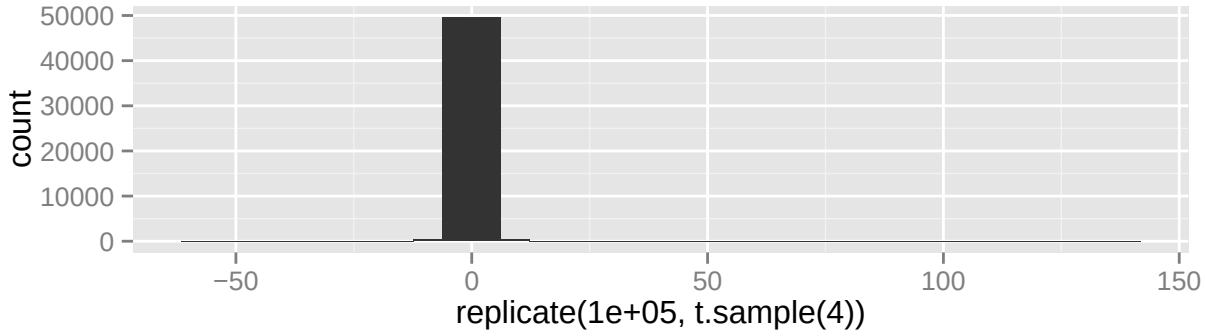
```
z.sample = function(n) {
    x = rnorm(n, 0, 1)
    return(mean(x)/1 * sqrt(n))
}

t.sample = function(n) {
    x = rnorm(n, 0, 1)
    return(mean(x)/sd(x) * sqrt(n))
}

qplot(replicate(1e+05, z.sample(4)))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(replicate(1e+05, t.sample(4)))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Examples and Excercises

Powerball Examples

Expected Value of Powerball lottery. - Five white balls drawn without replacement from 59, order irrelevant - red ball drawn from bin of 35 - $2 to play - assume no other players

What is the expected value

```r
#probability of getting 5 white balls
#five factorial 5 chances for first value, then 4 chances for second value, then..
#out of 59 chances for first value, 58 chances for second value.
white <- factorial(5) / prod((59-5):59)
red <- 1/ 35
Prob <- white*red

#does it work for simplier example
#deck of 5 cards, draw three. need to get the three target values 1,2,3

#first draw could be any of the three
d1 <- 3/5
#second draw the same.
d2 <- 2/4
#third draw the same
d3 <- 1/3
p  <- d1*d2*d3
p
```

```
## [1] 0.1
```

```r
#this gives us correct answer- check with permuations function.

permutations(n=5,r=3,repeats.allowed=F) %>%
  data.frame %>%
  summarise(win = mean(X1+X2+X3==6))
```

```
##   win
## 1 0.1
```