Unit 5: sampling distributions

Stats on stats on...

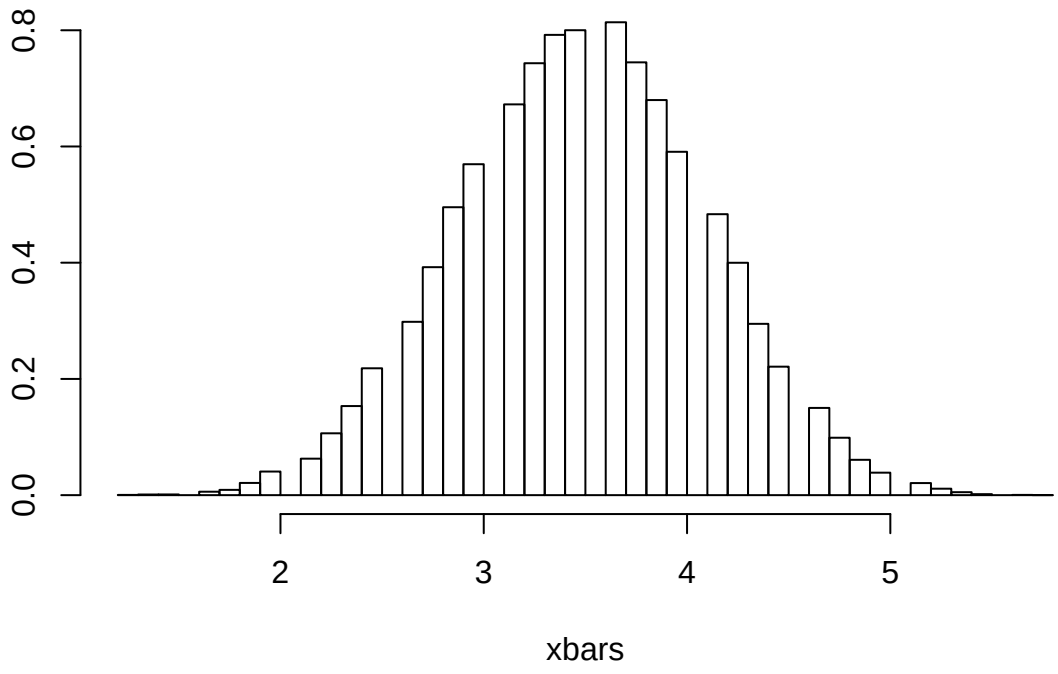
Let's return to the die throwing example, this time we will plot histograms of how the sample mean is distributed

The sample mean of independent observations is *normally* distributed about the population mean due to CLT - no matter how the underlying data is distributed!

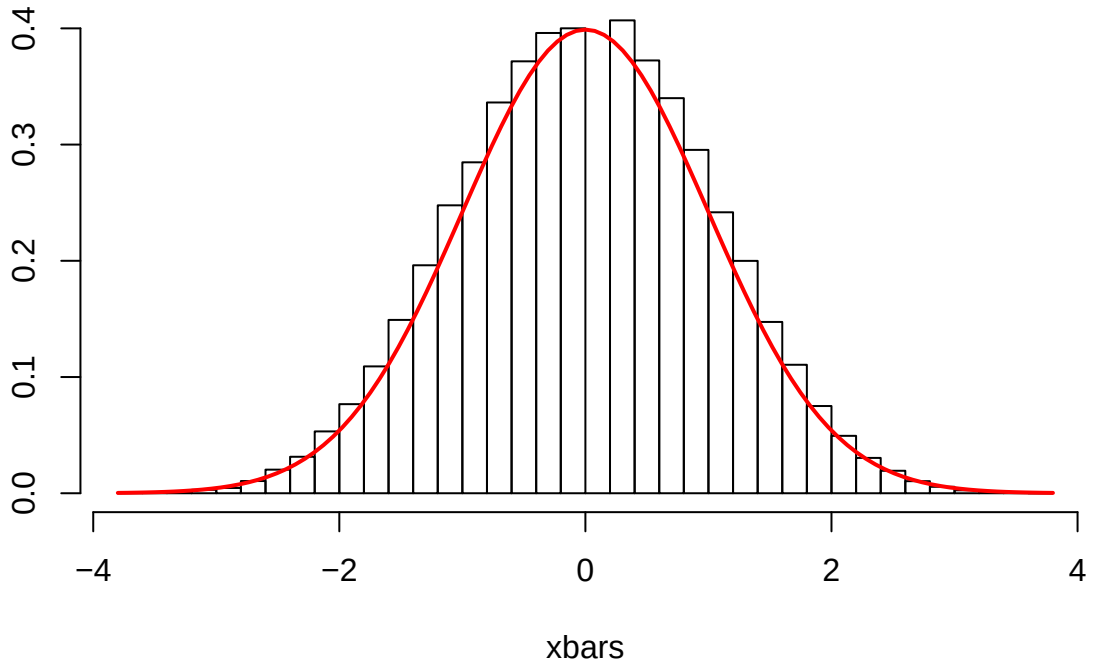
This code below returns m means each of sample size n dice throws

```
get_mean_dist <- function(m,n) {
  die <- 1:6
  replicate(m,mean(replicate(n,sample(die,size=1,replace=TRUE))))
}
```

```
xbars <- get_mean_dist(100000,8)
hist(xbars,freq = F,breaks=50,prob=TRUE,ylab=NULL,main=NULL)
```



$\frac{\overline{X}-3.5}{\frac{1.7}{\sqrt{n}}}$ should be $\sim N(0,1)$ (red curve) which it is!!



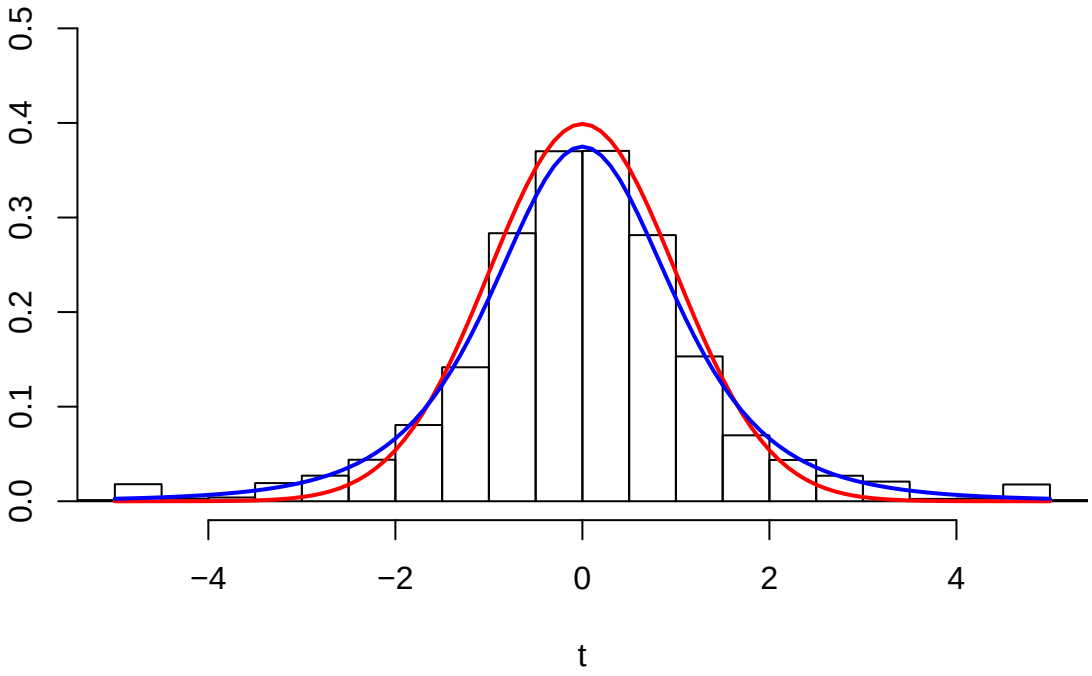
What if we didn't know the variance?

Consider

$t = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$

This, quantity will be collected on each step using a similar code to before.

```
get_t <- function(n) {
  die <- 1:6
  rolls <- replicate(n,sample(die,size=1,replace=TRUE))
 (mean(rolls) - 3.5)/(sd(rolls)/sqrt(n))   # this is t
}
  t_values <- replicate(500000,get_t(5)) #
```



Notice, it's very subtle but there is a difference as to what's going on in the tails!! The blue curve is student t dist with $\nu = n - 1 = 5 - 1 = 4$ dfs. Notice it fits the data better than the red curve which is a normal distribution $N(0, 1)$

t-distribution

The quantity t is the difference between the population mean and our sample mean divide by our estimate of the standard error. It turns out the distribution function for this quantity was calculated by William Gossett in 1908. The t-distribution is like the normal distribution but it has a little bit fatter tails.

More of the probability goes into the tails because our estimate of the population mean from the sample mean will be off by an extra factor due to our lack of knowledge of the population variance. As the sample size $n \to \infty$ the t distibution approaches the normal since $S \to \sigma$ in this same limit

The real power of the t distribution is when it is used for the *t- test* where we don't know either the population mean μ or the population standard deviation σ and want to estimate μ from a sample mean.

t - probability density

One of the nice things about taking a computational approach to the t distribution is that you do not need to derive formulas like the one below which is the probability density:

$$P(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

This formula describes the blue curve, on the previous slide with $\nu = 4\ldots$ but we don't care about the formulas...

We just need to know how much probability is in the tails... we can just sample it...

χ^2 distribution

Another important statistic that will arise later is the χ^2-statistic, which is just the sum of the squares of k independent standard normally distributed variates.

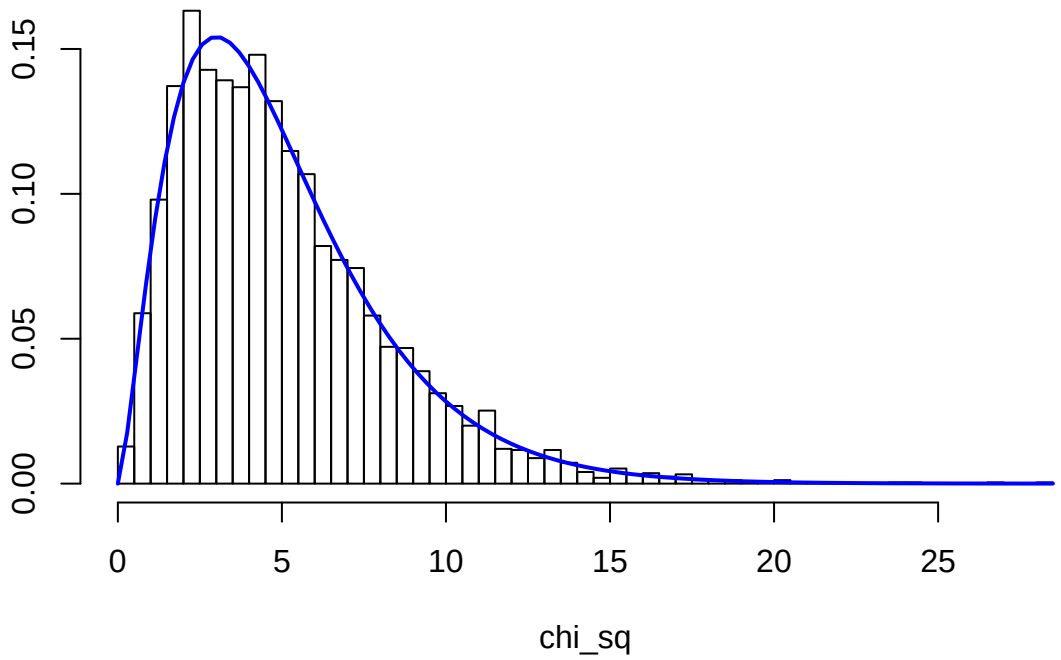
$\chi^2_k \sim \sum_{i=1}^{k} Z_i^2$ where each $Z_i \sim N(0,1)$

where the symbol $\sim$ is to be read "is distributed as".

To find out where the probability is, we again take the computational approach!

```
get_chi_sq <- function(k) {
  Z <- rnorm(k)
 chi_sq <- sum(Z%*%Z) # this is t
}
  chi_sq <- replicate(5000,get_chi_sq(5))
```

```
hist(chi_sq,freq = F,breaks=50,prob=TRUE,ylab=NULL,main=NULL)
curve(dchisq(x,5),
      col="blue", lwd=2, add=TRUE, yaxt="n")
```



Note that the actual formula for the probability density (blue curve on previous slide) is

$$P(x;\,k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

where $k = 5$

Summary Unit 5

We see that certain quantities can be described by "probabilities" which characterize how likely certain values are to occur. We shall see that these distributions will play a pivotal role in testing for the significance – as when we test for significance, we usually mean that something is significant if it only happens with a small probability - thus **significance testing is concerned with the unlikely tail events of distributions**

Lab: Unit 5

Exercise 1 - Use either python or R, your choice Write a code that generates a normalized histogram of the sample mean statistic

$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

drawn from m, n-sized samples from a exponential distribution where $\mu \equiv 1/\lambda = 12$ and $\sigma = 1/\lambda^2$. Show that this statistic is standard normally distributed $\sim N(0,1)$ by superimposing the probability density function onto the same figure as your histogram. m should be ≥ 1000. n can be any number more than 2. This will require you to be able to create normalized histograms in python or can also try to do this in R. In R, You can use `hist()` with the `prob=TRUE` argument to keep the area under the histogram curve approximately equal to 1

Exercise 2 - Use either python or R, your choice

Write a code that generates the t distribution as a function of n by sampling from the sample means over m, n sized samples from an exponential distribution with $\mu = 1/\lambda = 7$. Check that the tails of the t distribution that you create by sampling are fatter than that of a standard normal. I suggest using $n = 6$, and $m = 100000$

You will need to collect m observations of the sample statistic

$t = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$

This will require you to be able to create normalized histograms in python or can also try to do this in R. In R, You can use `hist()` with the `prob=TRUE` argument to keep the area under the histogram curve approximately equal to 1