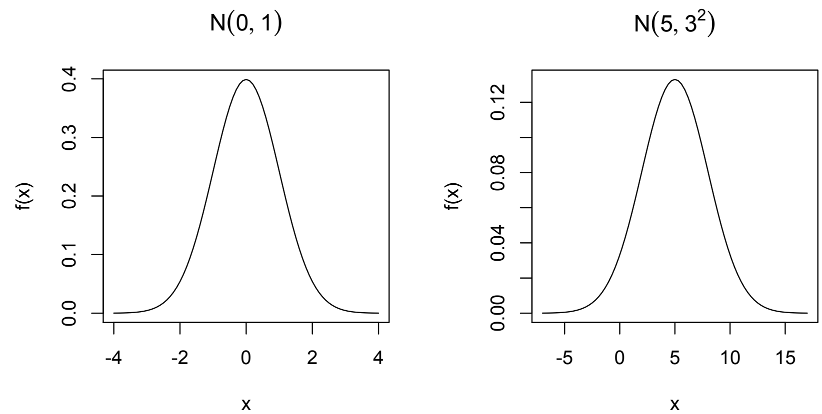
STA 250: Statistics

Lab 4

This lab overviews graphical and numerical summarization of pdfs and pmfs of scalar variables.

Plotting a pdf or a pmf over an appropriate range

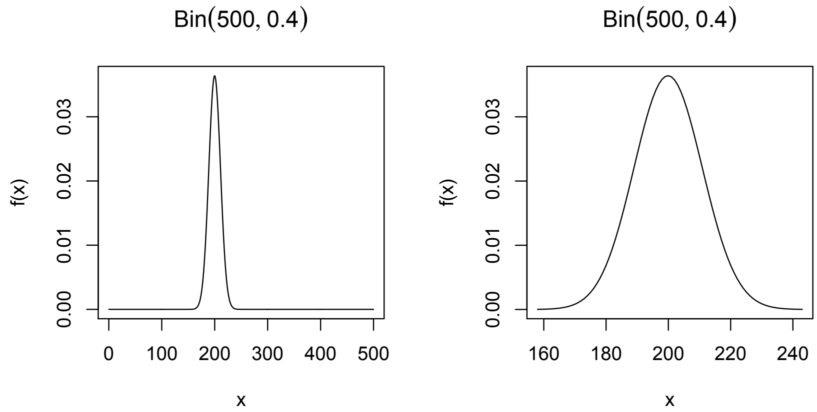
This is a trivial task if you have access to a `R` function `f()` that returns the pdf/pmf value `f(x)` at any input `x`. The only concern is to choose an appropriate range for x to display the curve. To display a $\mathsf{Normal}(0,1)$ pdf we could use a range $[-4,4]$, because most of the area under the $\mathsf{Normal}(0,1)$ bell curve is within this range (more than 0.9999). For an arbitrary $\mathsf{Normal}(\mu,\sigma^2)$ pdf, we can display the range $\mu \mp 4\sigma$, i.e., mean plus minus 4 standard deviations, which contains the same area under the $\mathsf{Normal}(\mu,\sigma^2)$ curve as does [-4,4] for the $\mathsf{Normal}(0,1)$ bell curve:



> Task 1. Make a *nice* plot of the $\mathsf{Normal}(-500,40^2)$ pdf. To annotate the axes and add a plot title, use:
> `plot(..., xlab = "x", ylab = "f(x)", main = expression(N(-500, 40^2)))`.

The same strategy applies while plotting other pdfs or pmfs. The $\mathsf{Binomial}(500,0.4)$ pmf is supported on the points $0,1,\cdots,500$. But it tightly concentrats around its mean $500\times 0.4 = 200$, and so plotting it over the whole range is rather useless (see the left panel of the Figure below). Instead, we can find a range $[a,b]$, such that pmf puts only a small probability, say 10^{-4}, outside this range. We can find the two end points so that the left out probability is equally split at the two tails. Then we must have a as the $10^{-4}/2$-th quantile and b as the $(1-10^{-4}/2)$-th quantile of the $\mathsf{Binomial}(500,0.4)$ distribution. The following code generates the right panel of the Figure below.

```
a <- qbinom(1e-4 / 2, 500, 0.4)
b <- qbinom(1e-4 / 2, 500, 0.4, lower = FALSE)
x <- a:b
plot(x, dbinom(x, 500, 0.4), ty = "l", ann = FALSE)
title(xlab = "x", ylab = "f(x)", main = expression(Bin(500, 0.4)))
```



TASK 2. What's the role of "`lower = FALSE`" in the second `qbinom()` call above?

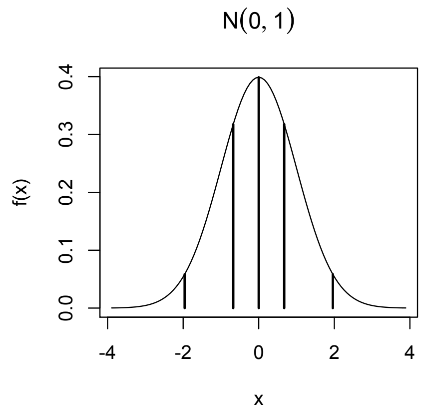
TASK 3. Make plots for pdf/pmf of the Poisson(30), Gamma(2, 1/5), Logis(4, 10) and t_4 distributions.

Quantiles of a distribution

The quantiles of a distribution offer an excellent way to summarize the distribution numerically or graphically. We already saw an example of this above in choosing the range of x values $[a, b]$ on which we plot the pdf/pmf. While those give a sense of the extreme, we can also use quantiles to mark the central parts of the distribution.

Recall that for any $u \in (0, 1)$, the u-th quantile of a probability distribution with cdf $F(x)$ is defined to be the point x_u such that $F(x_u) = u$. That is a vertical slice at $x = x_u$ splits the total pdf area/pmf total of 1 into u on the left and $1 - u$ on the right. The 0.5-th quantile is called the median of the distribution, it splits the pdf area into half and half. The 0.25-th and the 0.75-th quantiles are called the first and the third quartiles (median being the second quartile). Between the 0.025-th and the 0.975-th quantile, the pdf/pmf packs 95% of its area. These five quantiles together give an excellent annotation of a pdf/pmf. The following code and plot show these for Normal(0, 1).

```
qts <- qnorm(c(.025, .25, .5, .75, .975), 0, 1)
a <- qnorm(1e-4/2, 0, 1)
b <- qnorm(1e-4/2, 0, 1, lower = FALSE)
x <- seq(a, b, length = 101)
plot(x, dnorm(x, 0, 1), ty = "l", ann = FALSE)
title(xlab = "x", ylab = "f(x)", main = expression(N(0,1)))
segments(qts, 0 * qts, qts, dnorm(qts, 0, 1), lwd = 2)
```



TASK 4. Repeat task 3 now with the five-quantile summaries overlaid on each pdf/pmf.

Getting quantiles from samples

When R does not provide a quantile function for a distribution, we can still approximate its quantiles by drawing random samples from it. The code below compares a direct calculation against a sample based approximation of quantiles for $\mathsf{Normal}(0, 1)$

```
> qnorm(c(.025, .25, .5, .75, .975), 0, 1)
[1] -1.96 -0.67  0.00  0.67  1.96
> as.numeric(quantile(rnorm(1e4), c(.025, .25, .5, .75, .975)))
[1] -1.995 -0.690 -0.015  0.656  1.900
```

TASK 5. Get five-quantile summaries of $\mathsf{Poisson}(30)$, $\mathsf{Gamma}(2, 1/5)$, $\mathsf{Logis}(4, 10)$ and t_4 based on random samples drawn from these distributions (use 10^4 draws) and compare them with the actual values.

Consider the model $X \sim \mathsf{Binomial}(20, p)$ with prior $\xi(p) = \mathsf{Beta}(10, 10)$. In lecture we discussed that this defines a joint pd/mf $h(x, p)$ for (X, p). From the joint we could extract the marginal pmf for X. While it is possible to write down a formula for this marginal pmf, it is much easier to draw samples of X from this marginal pmf through the following:

```
p <- rbeta(1, 10, 10)
x <- rbinom(1, 20, p) ## using the ‘p’ generated above
```

which you can wrap inside `replicate()` to generate many samples at one go.

TASK 6. Use the above code with `replicate()` to generate 10^4 samples of X from its marginal and get their five-quantile summary.