Michael Leibert

Math 640

Homework 1

1. Show that the conditional distribution is a valid pdf/pmf for both discrete and continuous random variables. State the assumptions necessary to show this.

I am referencing Casella & Berger, pages 36, 148-150.

A function $f_Z(z)$ is a pdf (or pmf) of a random variable Z iff

a. $f_Z(z) \geq 0 \; \forall \; z$

b. $\displaystyle\sum_z f_Z(z) = 1$ (pmf) or $\displaystyle\int_{-\infty}^{\infty} f_Z(z) \, \mathrm{d}z = 1$ (pdf).

Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$ the conditional pmf of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y | X = x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $P(Y = y) = f_Y(y) > 0$ the conditional pmf of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = P(X = x | Y = y) = \frac{f(x, y)}{f_Y(y)}.$$

Since we have called $f(y|x)$ a pmf, we should verify that this function of y does indeed define a pmf for a rnadom variable. First, $f(y|x) \geq 0$ for every y since $f(x, y) \geq 0$ and $f_X(x) \geq 0$. Second,

$$\sum_y f(y|x) = \frac{\sum_y f(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

Thus, $f(y|x)$ is indeed a pmf.

Continuous:

Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$ the conditional pdf of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $P(Y = y) = f_Y(y) > 0$ the conditional pmf of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

Since we have called $f(y|x)$ a pdf, we should verify that this function of y does indeed define a pdf for a random variable. First, $f(y|x) > 0$ for every y since $f(x, y) > 0$ and $f_X(x) > 0$. Second,

$$\int_{-\infty}^{\infty} f(y|x) \, \mathrm{d}y = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_X(x)} \, \mathrm{d}y = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}y = \frac{f_X(x)}{f_X(x)} = 1.$$

Thus, $f(y|x)$ is indeed a pdf.

2. For each of the random variables, use a kernel recognition approach to determine the expectation.

$\theta \sim Beta(\alpha, \beta)$

$$E[\theta] = \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \, \theta \, \mathrm{d}\theta$$

$$= \frac{\Gamma(\alpha + \beta) \, \Gamma(\alpha + 1)}{\Gamma(\alpha + \beta + 1) \, \Gamma(\alpha)} \underbrace{\int_0^1 \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + 1)\Gamma(\beta)} \theta^{\alpha}(1 - \theta)^{\beta-1} \, \mathrm{d}\theta}_{Beta(\alpha+1,\beta)}$$

$$= \frac{\cancel{\Gamma(\alpha + \beta)} \, \alpha \, \cancel{\Gamma(\alpha)}}{(\alpha + \beta) \, \cancel{\Gamma(\alpha + \beta)} \, \cancel{\Gamma(\alpha)}} \cdot 1$$

$$= \frac{\alpha}{\alpha + \beta}$$

$\mu \sim Pois(\lambda)$

$$E[\mu] = \sum_{\mu=0}^{\infty} \frac{\lambda^{\mu}\ \exp(-\lambda)}{\mu!}\ \mu$$

$$= \sum_{\mu=1}^{\infty} \frac{\lambda^{\mu} \exp(-\lambda)}{(\mu-1)!}\ \frac{\lambda}{\lambda}$$

$$= \lambda \sum_{\mu=1}^{\infty} \frac{\lambda^{\mu-1}\ \exp(-\lambda)}{(\mu-1)!} \qquad \text{Let } \theta = \mu - 1$$

$$= \lambda \sum_{\theta+1=1}^{\infty} \frac{\lambda^{\theta}\ \exp(-\lambda)}{\theta!}$$

$$= \lambda \underbrace{\sum_{\theta=0}^{\infty} \frac{\lambda^{\theta}\ \exp(-\lambda)}{\theta!}}_{\text{Pois}(\lambda)}$$

$$= \lambda \cdot 1$$

$$= \lambda$$

$\nu \sim Binom(n,p)$

$$E[\nu] = \sum_{\nu=0}^{\infty} \binom{n}{\nu}\ p^{\nu}(1-p)^{n-\nu}\ \nu$$

$$= np \sum_{\nu=1}^{\infty} \frac{(n-1)!}{(\nu-1)!(n-\nu)!}\ p^{\nu-1}(1-p)^{n-\nu} \qquad \text{Let } \theta = \nu - 1$$

$$= np \sum_{\theta+1=1}^{\infty} \frac{(n-1)!}{\theta!(n-\theta-1)!}\ p^{\theta}(1-p)^{n-1-\theta}$$

$$= np \underbrace{\sum_{\theta=0}^{\infty} \binom{n-1}{\theta} p^{\theta}(1-p)^{n-1-\theta}}_{Binom(n-1,p)}$$

$$= np \cdot 1$$

$$= np$$

$\gamma \sim Exp(\beta)$

$$E[\gamma] = \int_{0}^{\infty} \beta \exp\left(-\beta\gamma\right) \gamma\ \mathrm{d}\gamma$$

$$= \frac{\beta}{\beta^2}\ \Gamma(2) \underbrace{\int_{0}^{\infty} \gamma^{2-1} \exp\left(-\beta\gamma\right) \frac{\beta^2}{\Gamma(2)}\ \mathrm{d}\gamma}_{Gamma(2,\beta)}$$

$$= \frac{1}{\beta} \cdot 1 \cdot 1$$

$$= \frac{1}{\beta}$$

3. Returning to our example from class, suppose we wish to study the number of ponderosa pine trees in the Black Hills of South Dakota that are infected with the mountain pine beetle (a beetle that ultimately kills ponderosapines). Let X be the number of infected ponderosa pines in a 10 square mile plot of forested land in the Black Hills. Further, let N be the number of ponderosa pines in that plot and P be the now *unknown* probability that a randomly selected pine is infected. Thus, $X|N, P \sim Binom(N, P)$ where N and P are independent random variables. For N, we assume $N|\Lambda \sim Pois(\Lambda)$ where $\Lambda \sim Gamma(\alpha, \beta)$. For P, we assume $P \sim Beta(\gamma, \zeta)$. Find $E(X)$.

$$
\begin{aligned}
E[X] &= E\Big[E(X|N, P)\Big] \\
&= E[NP] \\
&= E[N]\, E[P] \\
&= E\Big[E(N|\Lambda)\Big]\, \frac{\gamma}{\gamma + \zeta} \\
&= E\Big[\Lambda\Big]\, \frac{\gamma}{\gamma + \zeta} \\
&= \frac{\alpha}{\beta}\, \frac{\gamma}{\gamma + \zeta}
\end{aligned}
$$

4. Consider now an investigation of changes in weight depending on two different weight loss strategies: exercise only or diet only. Let β denote the average baseline weight, ϵ denote the average weight change after five months of those who were on the exercise only strategy, and δ denote the average weight change after five months of those on the diet only strategy. Further assume subjects are randomly assigned to either treatment arm with probability 0.5 and let T_i be 0 if the subject exercised only and 1 if the subject dieted only. For a randomly subject i, we then assume

$$
Y_i|\beta, \epsilon, \delta, T_i \sim N\left[\beta + \epsilon \cdot 1(T_i = 0) + \delta \cdot 1(T_i = 1), \sigma^2\right]
$$

Since β is average baseline weight, it cannot be negative nor can it be too small. Thus we assume $\beta \sim Gamma(1800, 10)$. Further, assume $\epsilon \sim N(e, \tau_e^2)$ and $\delta \sim N(d, \tau_d^2)$. Assuming that T_i is independent of ϵ and δ, find $E(Y_i)$, the expected weight at study end for subject i.

$$
\begin{aligned}
E[Y_i] &= E\Big[E(Y_i|\beta, \epsilon, \delta, T_i)\Big] \\
&= E\Big[\beta + \epsilon \cdot 1(T_i = 0) + \delta \cdot 1(T_i = 1)\Big] \\
&= E\Big[\beta\Big] + E\Big[\epsilon \cdot 1(T_i = 0)\Big] + E\Big[\delta \cdot 1(T_i = 1)\Big] \\
&= 180 + E\Big[\epsilon\Big]\, E\Big[1(T_i = 0)\Big] + E\Big[\delta\Big]\, E\Big[1(T_i = 1)\Big] \\
&= 180 + 0.5e + 0.5d
\end{aligned}
$$

5. Use the Probability Integral Transform to write code that generates samples from each of the following distributions (you may look up the CDFs of each, i.e. you do not have to derive them). Generate 1000 samples each and plot the empirical distribution overlaying the theoretical distribution. Then increase the number of samples to 10,000 and repeat. For reproducibility, please set the seed *before each run* to 108, `set.seed(108)`. Present all your code in your solution.

$\theta \sim Exp(27)$

$$F(x) = 1 - \exp(-27x)$$
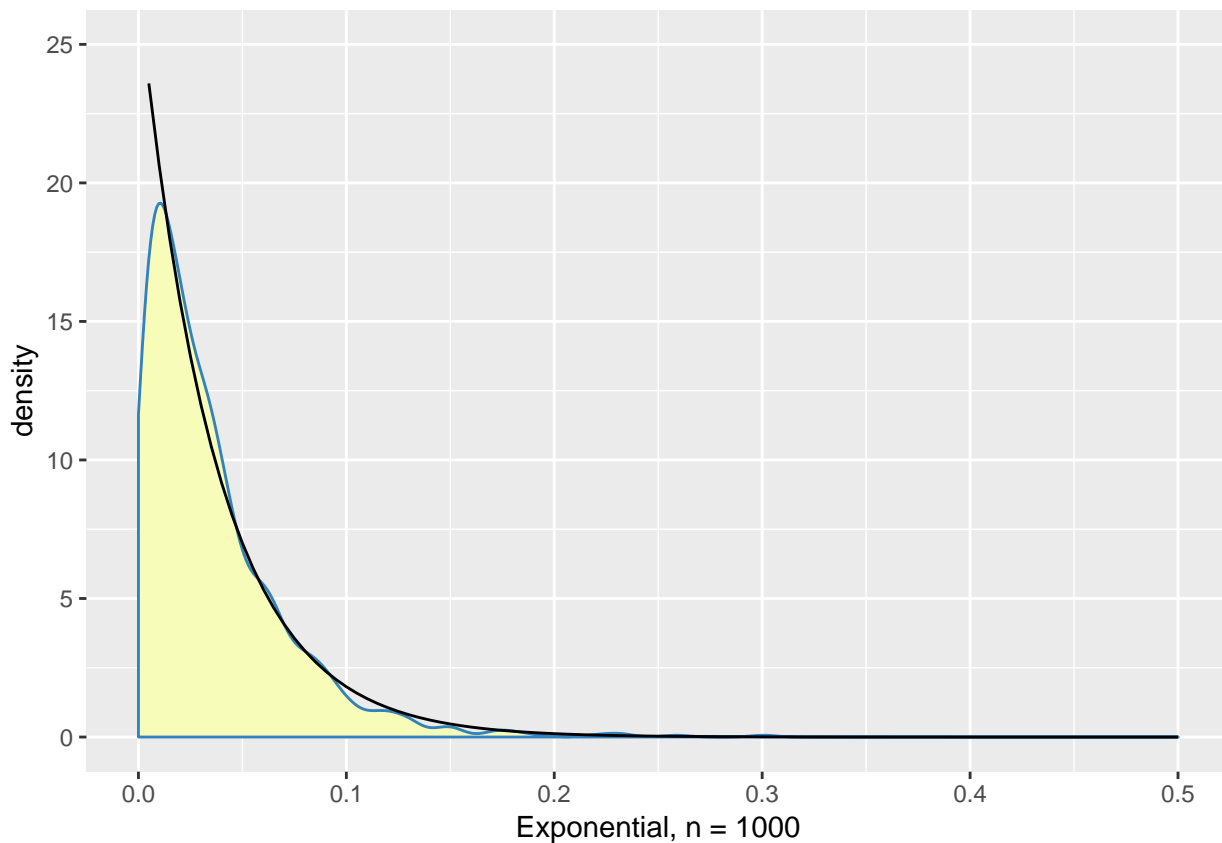$$u = 1 - \exp(-27x)$$
$$1 - u = \exp(-27x)$$
$$\frac{1}{1-u} = \exp(27x)$$
$$x = \frac{1}{27} \log\left(\frac{1}{1-u}\right)$$

```
set.seed(108)
n = 1000;    u <- runif(n)
x = (1/27) * log(1 / (1-u) );    dat<-data.frame(x)

ggplot(dat, aes(x)) +  geom_density( col="#3182bd",fill="#f7fcb9" ) +
    labs(x="Exponential, n = 1000") + xlim( 0, .5 ) + ylim(0,25) +
    stat_function( fun = dexp  , args = list(rate =  27) )
```
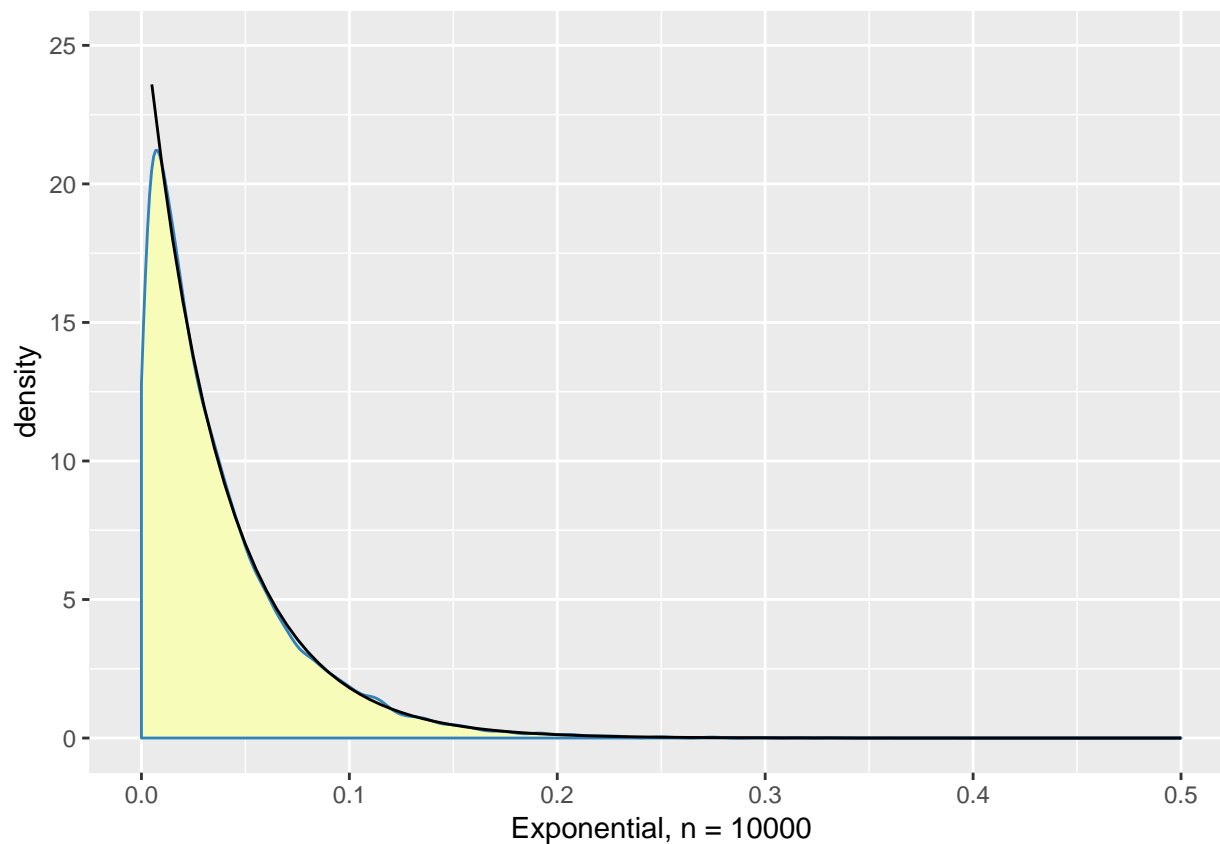
```
## Warning: Removed 1 rows containing missing values (geom_path).
```

```
set.seed(108)
n = 10000;     u <- runif(n)
x = (1/27) * log(1 / (1-u) );    dat<-data.frame(x)

ggplot(dat, aes(x)) +  geom_density( col="#3182bd",fill="#f7fcb9" ) +
    labs(x="Exponential, n = 10000") + xlim(0 , .5 ) + ylim(0,25) +
    stat_function( fun = dexp  , args = list(rate =  27) )
```

Warning: Removed 1 rows containing missing values (geom_path).

$\theta \sim Cauchy(-7, 2)$, a location of -7 and scale of 2

Note: the Cauchy sampler gives quite a few values that could be considered extreme relative to the rest of the distribution. In order to generate a decent looking plot, I have set the x-axis between -50 and 50, with the knowledge I have omitted many of the samples.

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x - \mu}{\sigma}\right)$$

$$u = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x + 7}{2}\right)$$

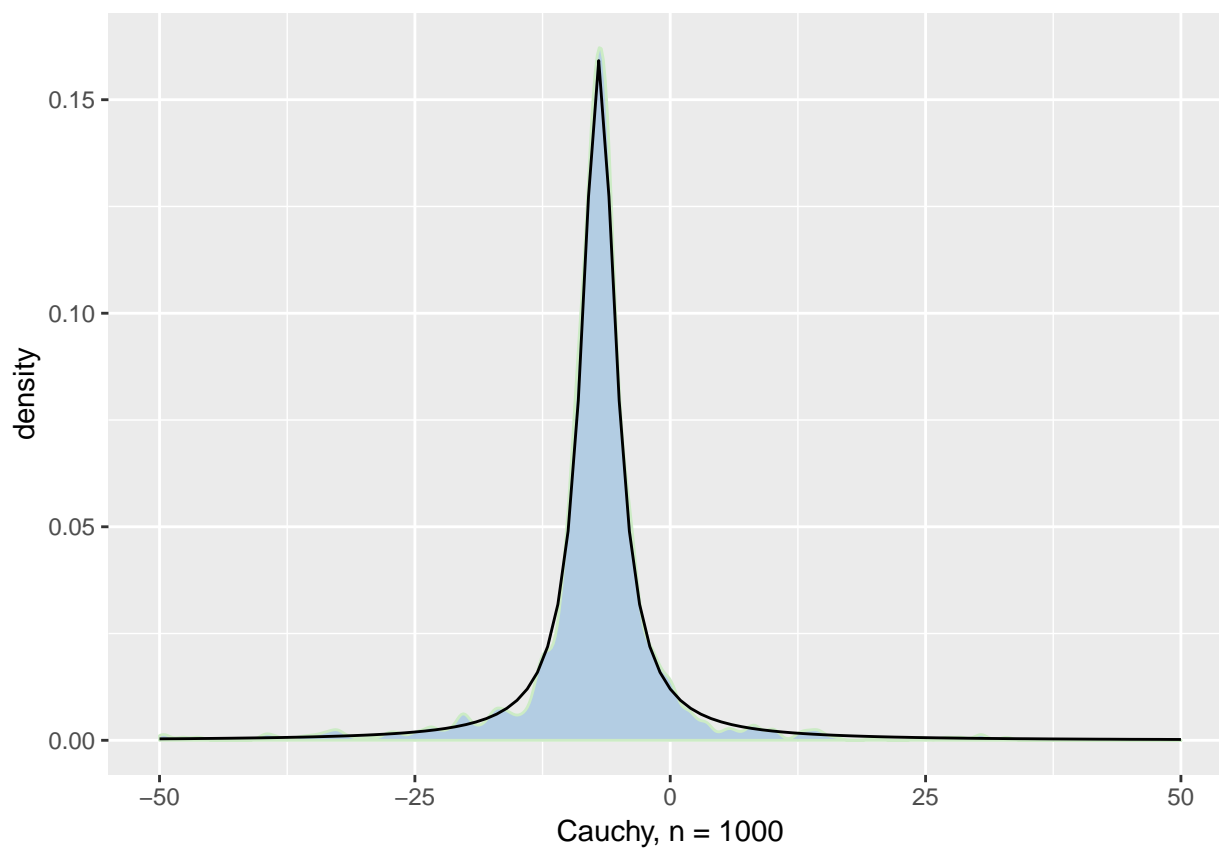$$\pi u - \frac{\pi}{2} = \arctan\left(\frac{x + 7}{2}\right)$$

$$\tan\left(\pi u - \frac{\pi}{2}\right) = \frac{x + 7}{2}$$

$$2 \tan\left(\pi u - \frac{\pi}{2}\right) - 7 = x$$

```
w <- 50
set.seed(108)
n = 1000;    u <- runif(n)
x = 2 * tan( (pi * u ) - ( pi / 2 ) ) - 7;    dat<-data.frame(x)

ggplot(dat, aes(x)) +  geom_density( col="#ccebc5",fill="#b3cde3" ) +
       labs(x="Cauchy, n = 1000")    +    xlim( -w , w ) +
    stat_function( fun = dcauchy , args = list( location = -7, scale = 2) )
```

```
## Warning: Removed 30 rows containing non-finite values (stat_density).
```
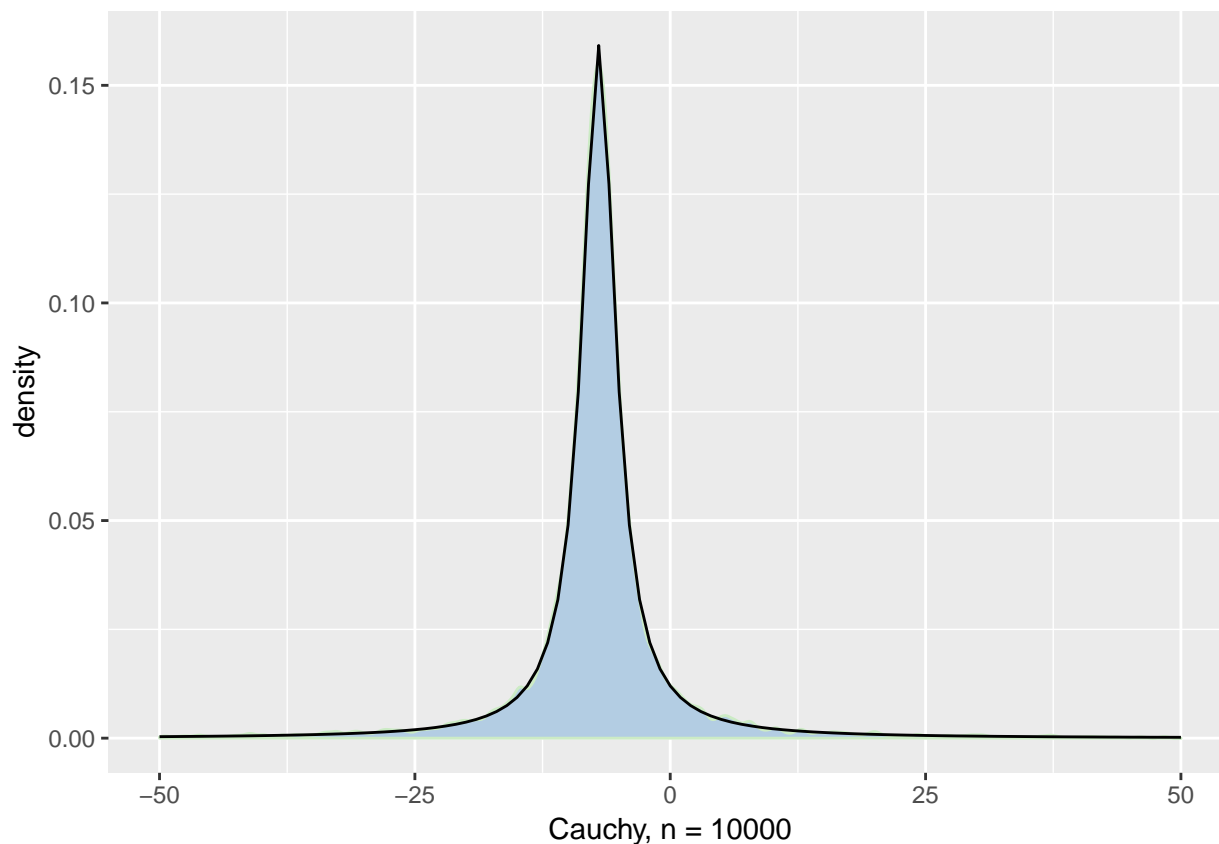
```
set.seed(108);  n = 10000;    u <- runif(n)
x = 2 * tan( (pi * u ) - ( pi / 2 ) ) - 7;   dat<-data.frame(x)

ggplot(dat, aes(x)) + geom_density( col="#ccebc5",fill="#b3cde3" ) +
       labs(x="Cauchy, n = 10000")    +    xlim( -w , w ) +
    stat_function( fun = dcauchy , args = list( location = -7, scale = 2) )
```

Warning: Removed 282 rows containing non-finite values (stat_density).

$\theta \sim Gumbel(3, 6)$, a location of 3 and scale of 6

$$F(x) = \exp\left[-\exp\left(\frac{-(x - \mu)}{\beta}\right)\right]$$

$$u = \exp\left[-\exp\left(\frac{-(x - 3)}{6}\right)\right]$$

$$-\log(u) = \exp\left(\frac{-(x - 3)}{6}\right)$$

$$\log\left[-\log(u)\right] = \frac{3 - x}{6}$$

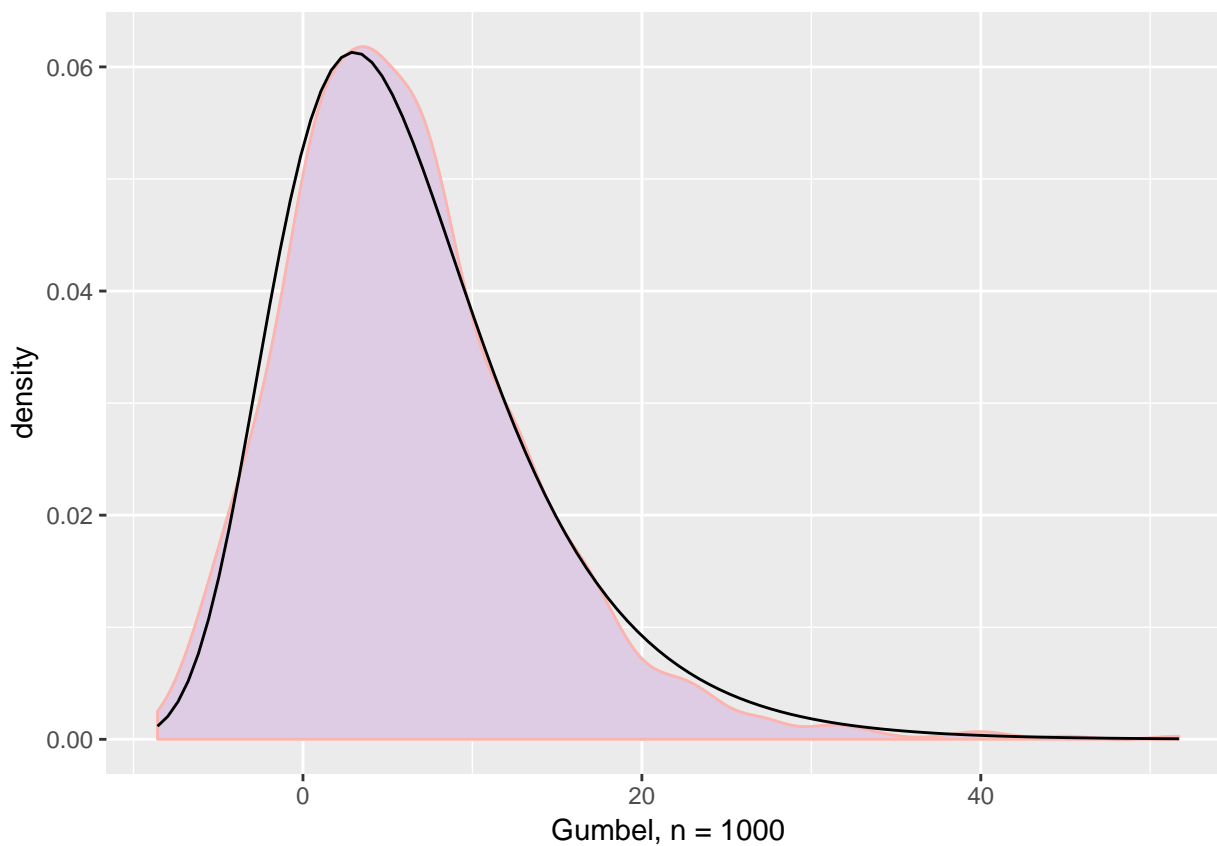$$3 - 6\log\left[-\log(u)\right] = x$$

```
set.seed(108)
n = 1000;    u <- runif(n)
x = 3 - 6*log( -log( u ) );    dat<-data.frame(x)

ggplot(dat, aes(x)) +  geom_density( col="#fbb4ae",fill="#decbe4" ) +
  labs(x="Gumbel, n = 1000") +
  stat_function( fun = dgumbel, args = list( loc = 3, scale = 6) )
```



```
set.seed(108)
n = 10000;    u <- runif(n)
x = 3 - 6*log( -log( u ) );    dat<-data.frame(x)

ggplot(dat, aes(x)) +  geom_density( col="#fbb4ae",fill="#decbe4" ) +
    labs(x="Gumbel, n = 10000")    +
    stat_function( fun = dgumbel, args = list( loc = 3, scale = 6) )
```

Gumbel, n = 10000