STAT 5034 Homework 1

Warren Geither

8/29/2020

Honor Code: "I have neither given nor received unauthorized assistance on this assignment."

Problem 1

a.) pmf

b.)

$$
\begin{aligned}
E(y) &= \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} \\
&= \sum_{y=1}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} \\
&= e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^y}{(y-1)!} \text{ , sub } x = y-1 \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x+1}}{(x)!} \\
&= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{(x)!} \text{ , using the identity } e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda \frac{e^{-\lambda}}{e^{\lambda}} \\
&= (\lambda)(1) \\
&= \lambda \ \blacksquare
\end{aligned}
$$

c.)

$$
\begin{aligned}
E(y^2) &= \sum_{y=0}^{\infty} y^2 \frac{\lambda^y e^{-\lambda}}{y!} \\
&= e^{-\lambda} \sum_{y=0}^{\infty} y \frac{\lambda^y}{(y-1)!} \text{ , the zero term is 0} \\
&= e^{-\lambda} \sum_{y=1}^{\infty} y \frac{\lambda^y}{(y-1)!} \text{ , sub } x = y-1 \\
&= e^{-\lambda} \sum_{x=0}^{\infty} (x+1) \frac{\lambda^{x+1}}{(x)!} \text{ , pulling out } \lambda \\
&= \lambda e^{-\lambda} \sum_{x=0}^{\infty} (x+1) \frac{\lambda^{x}}{(x)!} \text{ , distributing we get} \\
&= \lambda e^{-\lambda} (\sum_{x=0}^{\infty} (x) \frac{\lambda^x}{(x)!} + \sum_{x=0}^{\infty} (1) \frac{\lambda^x}{(x)!}) \text{ , using the identity } e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \\
&= \lambda e^{-\lambda} (\sum_{x=0}^{\infty} (x) \frac{\lambda^x}{(x)!} + e^{\lambda}) \text{ , redistributing } e^{-\lambda} \\
&= \lambda (\sum_{x=0}^{\infty} (x) \frac{e^{-\lambda}\lambda^x}{(x)!} + e^{-\lambda} e^{\lambda}) \text{ , recognizing we know have } E(x) \\
&= \lambda (\lambda + \frac{e^{\lambda}}{e^{\lambda}}) \\
&= \lambda (\lambda + 1) \\
&= (\lambda^2 + \lambda) \\
\implies Var(y) &= E(y^2) - (E(y))^2 \\
&= (\lambda^2 + \lambda) - (\lambda)^2 \\
&= \lambda \ \blacksquare
\end{aligned}
$$

d.)

- What is the probability that the robot makes 2 mistakes in an hour?

```
dpois(2, lambda = 2.3)
```

```
## [1] 0.2651846
```

```
(2.3^2)*(exp(-2.3))/2
```

```
## [1] 0.2651846
```

- What is the probability the robot makes at least 2 mistakes in an hour?

```
1 - sum(dpois(0:1, lambda = 2.3))
```

```
## [1] 0.6691458
```

```
1- (((2.3^0)*(exp(-2.3))/1) + ((2.3^1)*(exp(-2.3))/1))
```

```
## [1] 0.6691458
```

- What is the the probability that the robot makes 2.3 mistakes in an hour?
  - the domain is the Natural number union {0}, so we 2.3 isnt possible (i.e P(X=2.3)=0)

```
dpois(2.3, lambda = 2.3)
```

```
## Warning in dpois(2.3, lambda = 2.3): non-integer x = 2.300000
## [1] 0
```

- What is the probability that the robot makes 0 mistakes in an hour?

```
dpois(0, lambda = 2.3)
```

```
## [1] 0.1002588
```

```
((2.3^0)*(exp(-2.3))/1)
```

```
## [1] 0.1002588
```

- What is the probability that the robot makes less than 6 mistakes in an hour?

```
sum(dpois(0:5, lambda = 2.3))
```

```
## [1] 0.9700243
```

e.) The sample mean is a "good" estimator because it is the maximum likelihood estimator for lambda

```
y<-c(11,7,2,7,4,8,13,3,6,6,15,8,2,4,5,11,11,4,9,3,9,8,5,9,6)

est_lambda = sum(y)/25

print(est_lambda)
```

```
## [1] 7.04
```

Problem 2

a.)

$$g(y) = k(1+y)^{-k-1}$$

$$\begin{aligned}
&\lim_{n\to\infty} \int_0^n k(1+y)^{-k-1}dy \\
&= \lim_{n\to\infty} \frac{k}{-k}(1+y)^{-k}\Big|_0^n dy \\
&= \lim_{n\to\infty} -1(1+y)^{-k}\Big|_0^n dy \\
&= \lim_{n\to\infty} [-1(1+n)^{-k}] - [-1(1+0)^{-k}] \\
&= \lim_{n\to\infty} [\frac{-1}{(1+n)^k}] + [\frac{1}{1^k}] \\
&= [0] + [1] \\
&= 1
\end{aligned}$$

Thus g(y) is a pdf.

b.)

$$g(y) = ky^{-k-1}; y >= 1$$

$$\begin{aligned}
&\lim_{n\to\infty} \int_1^n ky^{-k-1}dy \\
&= \lim_{n\to\infty} \frac{k}{-k}(y)^{-k}\Big|_1^n dy \\
&= \lim_{n\to\infty} -1(y)^{-k}\Big|_1^n dy \\
&= \lim_{n\to\infty} [-1(n)^{-k}] - [-1(1)^{-k}] \\
&= \lim_{n\to\infty} [\frac{-1}{(n)^k}] + [\frac{1}{1^k}] \\
&= [0] + [1] \\
&= 1
\end{aligned}$$

Thus g(y) is a pdf.

Problem 3

a.) The temperature is between 37 and 65 with a mean of 54.10. There is 1 outlier value for the percent butterfat .446% but the first quartile starting at 4.6% This could be do to some kind of measurement error and we may want to remove the data for further analysis. Looking at the scatterplot, we can see what appears to be a small negative correlation between percent butterfat and temperature.

```
# temperature vector
x <- c(64,65,65,64,61,55,39,41,46,59,56,56,62,37,37,45,57,58,60,55)

# percent of butterfat
y <- c(4.65,4.58,4.67,4.60,4.83,4.55,5.14,4.71,4.69,4.65,4.36,4.82,4.65,4.66,4.95,4.60,4.68,4.65,4.6,.4

# bring in ggplot
library("ggplot2")

# bring in knitr
library("knitr")

# create dataframe
df = data.frame("temp" = x, "pfb" = y)

# Check out what the data looks like
kable(head(df,10))
```

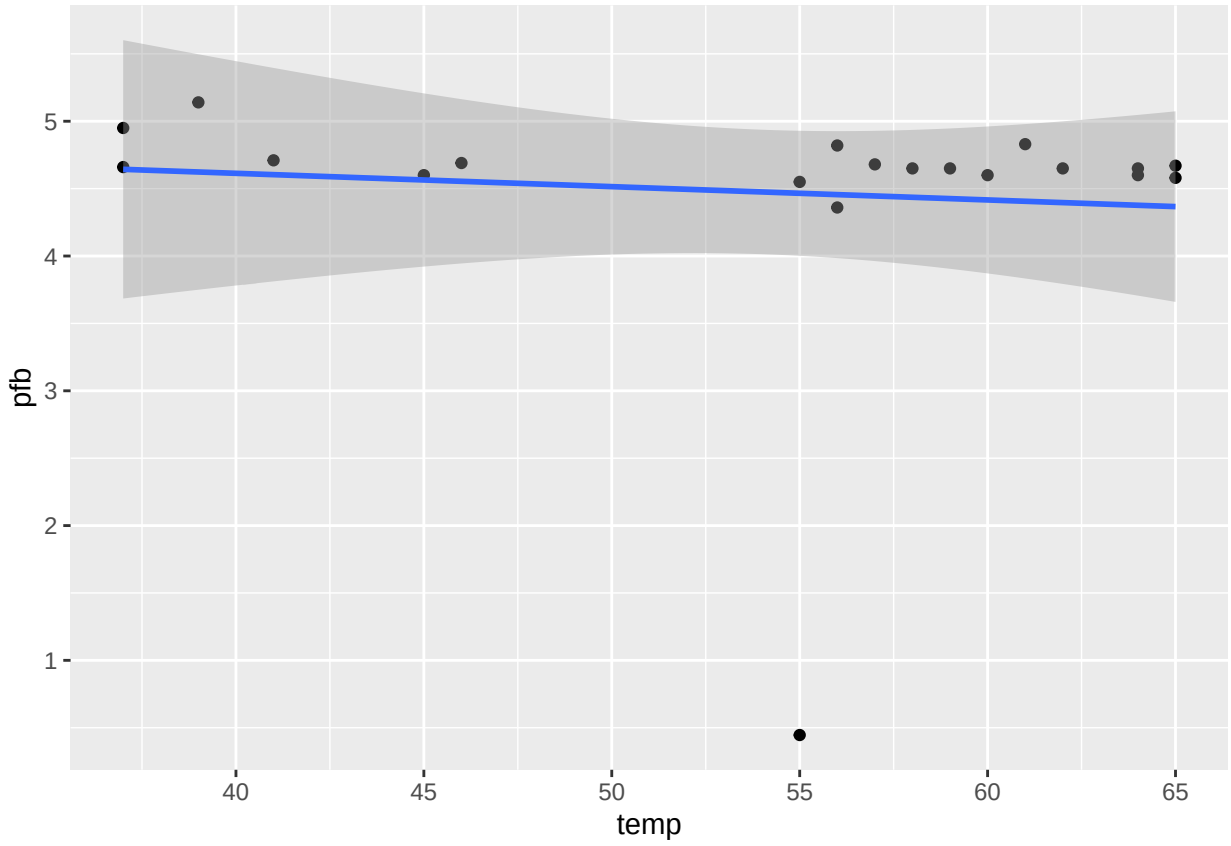
temp	pfb
64	4.65
65	4.58
65	4.67
64	4.60
61	4.83
55	4.55
39	5.14
41	4.71
46	4.69
59	4.65

```r
# run summary stats
kable(summary(df))
```

temp	pfb
Min. :37.00	Min. :0.446
1st Qu.:45.75	1st Qu.:4.600
Median :56.50	Median :4.650
Mean :54.10	Mean :4.474
3rd Qu.:61.25	3rd Qu.:4.695
Max. :65.00	Max. :5.140

```r
# scatterplot for temp vs. pbf
ggplot(data = df, aes(x=temp,y=pfb)) +
  geom_point() +
  geom_smooth(method=lm)
```

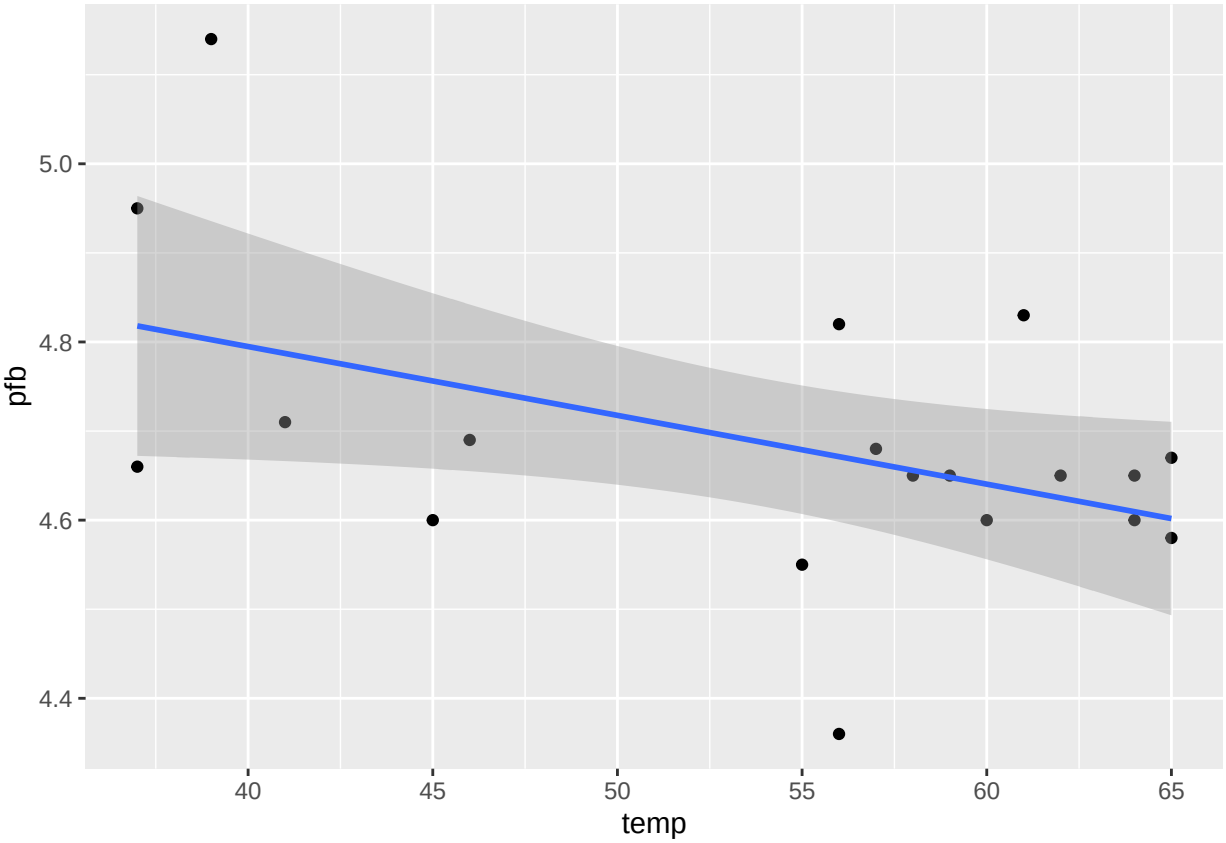
```
## `geom_smooth()` using formula 'y ~ x'
```




```r
# remove outlier
df_outlier_removed = df[-c(20), ]

# scatterplot for temp vs. pbf with outlier removed
ggplot(data = df_outlier_removed, aes(x=temp,y=pfb)) +
  geom_point() +
  geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



b.) Pearson's correlation coefficient describes the strength and direction of a linear relationship between variables.

When the outlier is included, they are negatively correlated (as temperature rises, percent of butterfat goes down), however the relationship is not very strong since it is only approximately -0.1.

When the outlier is not included, they are still negatively correlated, however the relationship is much stronger with the coefficient being much closer to -1 at approximately -0.47

```
# calculate pearson's correlation coefficient
temp = df[1]
pfb = df[2]
cor(x=temp, y=pfb, method="pearson")
```

```
##              pfb
## temp -0.09911088
```

```
# check out what the coefficient is with removal of the outlier
temp_2 = df_outlier_removed[1]
pfb_2 = df_outlier_removed[2]
cor(x=temp_2, y=pfb_2, , method="pearson")
```

```
##             pfb
## temp -0.4688517
```

c.) The outlier concerned me as I am assuming it is some type of measurement error. I would need to ask Dr. Osborne more about the data to confirm this.

Problem 4

a.) Parameter. Since its the true "average" of the whole population.

b.) Statistic. Because its based on a sample, not all of her customers did the survey (more than likely)

c.) Parameter & Statistic. Since its the true number of people in the US and was also their sample.

Problem 5

(a) Continuous
(b) Continuous
(c) Ordinal
(d) Continuous
(e) Count
(f) Binary
(g) Nominal
(h) Continuous or ordinal (if its in ranges)
(i) Binary
(j) Count
(k) Continuous