# Homework #7

Winter 2020, STATS 509
Dino Bektesevic

## Problem 1: Sampling Distributions

A researcher plans to carry out an opinion survey regarding a yes/no question. Suppose that (unknown to the researcher) 89% of people in the target population answer 'yes'. We will use 1 to denote 'yes' and 0 to denote 'no'.

a. Let $X_1$ be the response from the first person the researcher asks. Before this person is asked, what is the distribution of $X_1$. What is $E[X_1]$ and $V(X_1)$?

   First draw is just a single Bernoulli experiment with $p = 0.89$ so:

$$P(X = X_1) = p^k(1-p)^{1-k} = p = 0.89$$
$$E[X_1] = p = 0.89$$
$$V[X_1] = p(1-p) = 0.0979$$

The researcher plans to obtain a random sample of size 25: $X_1, \ldots, X_{25}$.
(Suppose either that the researcher is sampling with replacement or that the population is so large that we may regard the samples as being taken with replacement.)

b. Write down the distribution of $X_i$ for $i = 1, \ldots, 25$; also write down $E[X_i]$, $V(X_i)$ and $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$.

   Each draw is just a single Bernoulli experiment with $p = 0.89$ because of i.i.d. so:

$$P(X_i = x_1) = p^k(1-p)^{1-k} = p = 0.89$$
$$E[X_i] = p = 0.89$$
$$V[X_i] = p(1-p) = 0.0979$$
$$Cov(X_i, X_j) = E[X_iX_j] - E[X_i]E[X_j] = 0$$

If it's the joint distribution we are after then:

$$
\begin{aligned}
P(X_1 = x_1 \dots X_{25} = x_{25}) &= P(X_1 = x_1) \dots P(X_{25} = x_{25}) \\
&= \prod_{i=1}^{25} P(X_i = x_i) = 0.89^{25} \\
&= 0.054 \\
E[P(X_i)] &= E\left[\prod_{i=1}^{25} P(X_i = x_i)\right] \\
&= \prod_{i=1}^{25} E[P(X_i = x_i)] = 0.89^{25} \\
&= 0.054 \\
V[P(X_i)] &= E[(X_1 \cdots X_{25})^2] - (E[X_1 \cdots X_{25}])^2 \\
&= E[X_1^2] \cdots E[X_{25}^2] - E[X_1]^2 \cdots E[X_{25}]^2 \\
&= \prod_{i=1}^{25} E[X_i^2] - \prod_{i=1}^{25} (E[X_i])^2 \\
&= \prod_{i=1}^{25} V[X_i] + E[X_i]^2 - \prod_{i=1}^{25} (E[X_i])^2 \\
&= (0.0979 + 0.89^2)^{25} - 0.89^{50} \\
&= 0.0513 \\
Cov(X_i, X_j) &= 0
\end{aligned}
$$

c. What is the distribution of $\bar{X}$? *Hint: see lecture notes.*

What is the mean and variance of this distribution?

Lecture 10 slide 5 tells us that $X_i \sim Bernoulli(p)$ then $n\bar{X} \sim Binomial(n, p)$ and slide 8 of the same lecture also tells us via the sample mean theorem that the expected value of sample matches the expected value of the random variable but the variance reduces by the number of samples:

$$
\begin{aligned}
P(25\bar{X}) &\sim Bin(25, 0.9) \\
E[\bar{X}] &= p = 0.89 \\
V[\bar{X}] &= \frac{V[X_i]}{n} = 0.003916
\end{aligned}
$$

d. Use R or Python, replicate the experiment $10,000$ times.

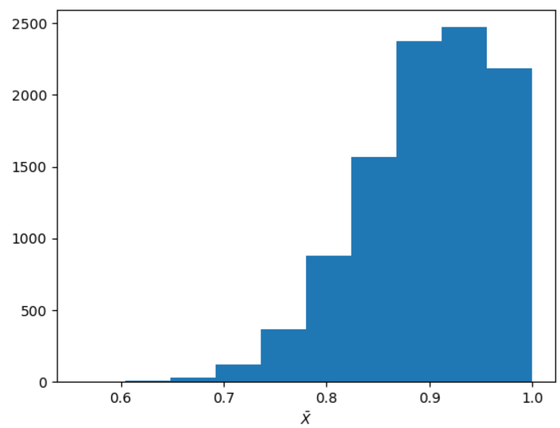
*Hint: Recall that a Bernoulli($p$) random variable is just a Binomial($1, p$) random variable; thus to obtain a sample of size 25 from a Bernoulli($0.89$) random variable, one might use:*

```
p <- 0.89

n <- 25

x <- rbinom(n,1,p)
```

*The vector* x *will contain the 25 Bernoulli($p$) observations, from which you can compute the sample mean. This should then be put in a loop which repeats this process, 10K times, storing the result.* Construct a histogram with the distribution of $\bar{X}$ for each sample. Does the distribution of sample means appear to be normal? Explain your answer.




The distribution does not look normal.

```python
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt


def problem1():
    n, p = 25, 0.89
    nTrials = 10000
    means = []
    for i in range(nTrials):
        x = stats.bernoulli.rvs(p, size=n)
        means.append(x.mean())

    means = np.array(means)
    print("Problem 1:")
    print(f"    Mean of means: {means.mean()}")
    print(f"    Variance of means: {means.std()**2}")
    print(f"    Mean squared error: {np.mean((means-p)**2)}")

    plt.hist(means)
    plt.xlabel(r"$\bar X$")
    plt.show()
```

e. Find the mean and variance of $\bar{X}$, based on your 10K simulations. Does this agree with your calculation in c.?

```
Mean of means: 0.8891480000000002

Variance of means: 0.003957434095999998
```

The simulated values match the calculated ones rather closely.

f Compute the mean squared error of $\bar{X}$ as an estimate of the unknown proportion $\theta = 0.89$ (i.e. find the squared error of the estimate, which is $(\bar{X} - \theta)^2$, in each repetition and then average over all repetitions). What do you notice? Give a simple explanation by referring to a result from the course.

```
Mean squared error: 0.0039581600000000005
```

It is very close to the variance of means calculated above.

## Problem 2

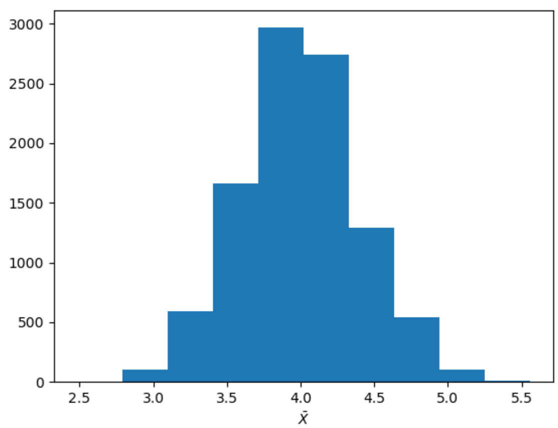
Given a sample $X_1, \ldots, X_n$ of independent Poisson($\lambda$) random variables, a researcher intends to use $\bar{X}$, the sample mean, as an estimate of $\lambda$. Suppose that $n = 20$, and $\lambda = 4$.

a. Write down the mean and variance of this distribution.

As in problem 1 c we have:

$$E[\bar{X}] = \lambda = 4$$
$$V[\bar{X}] = \frac{\lambda}{n} = 0.2$$

b. Using R or Python, replicate the experiment 10,000 times. Construct a histogram with the distribution of $\bar{X}$ for each sample. Does the distribution of sample means appear to be normal? Explain your answer.




The distribution looks normal.

```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
def problem2():
    n, lambd = 25, 4
    nTrials = 10000
    means = []
    for i in range(nTrials):
        x = stats.poisson.rvs(lambd, size=n)
        means.append(x.mean())

    means = np.array(means)
    print("Problem 2")
    print(f"    Mean of means: {means.mean()}")
    print(f"    Variance of means: {means.std()**2}")
    print(f"    Mean squared error: {np.mean((means-lambd)**2)}")

    plt.hist(means)
    plt.xlabel(r"$\bar X$")
    plt.show()
```

c. Find the mean and variance of $\bar{X}$, based on your 10K simulations. Does this agree with your calculation in a.?

```
Mean of means: 3.9952959999999997
```

```
Variance of means: 0.16090299238400002
```

d. Compute the mean squared error of $\bar{X}$ as an estimate of $\lambda$. Again what do you notice?

```
Mean squared error: 0.16092512
```

It is very close to the variance of means calculated above.

## Problem 3: Hypothesis Tests

Suppose that $Y_1, \ldots, Y_n$ are i.i.d samples from a Poisson distribution with parameter $\lambda$.

a. Find the log of the likelihood: $\ln f(y_1, \ldots, y_n \mid \lambda)$; Poisson distribution is given as

$$f(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

The likelihood can be written as:

$$L(\lambda|y_1 \ldots y_n) = \prod_{i=1}^{n} f(y_i|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}$$

Log likelihood is then:

$$l(\lambda|y_1 \ldots y_n) = \ln(\lambda)\sum_{i=1}^{n} y_i - n\lambda - \ln\prod_{i=1}^{n} y_i!$$

b. By differentiating the log likelihood, find the value $\hat{\lambda}$ of $\lambda$ that maximizes the likelihood; confirm that $\hat{\lambda}$ is a maximum. The estimate $\hat{\lambda}$ is called the *maximum likelihood estimator (MLE)*.

$$\begin{aligned}
\frac{d}{d\lambda} l(\lambda|y_1 \ldots y_n) &= \frac{d}{d\lambda}\ln(\lambda)\sum_{i=1}^{n} y_i - \frac{d}{d\lambda} n\lambda - \frac{d}{d\lambda}\ln\prod_{i=1}^{n} y_i! = 0 \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\hat{\lambda}} - n \\
n\hat{\lambda} &= \sum_{i=1}^{n} y_i \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} y_i}{\lambda} \\
\hat{\lambda} &= \bar{y}
\end{aligned}$$

To confirm it's a maximum we have to take a look at the second derivative:

$$\begin{aligned}
\frac{d^2}{d\lambda^2} l(\lambda|y_1 \ldots y_n) &= \frac{d}{d\lambda}\frac{\sum_{i=1}^{n} y_i}{\lambda} \\
&= -\frac{\sum_{i=1}^{n} y_i}{\lambda^2}
\end{aligned}$$

Since it's always negative this is a maximum.

Suppose that we wish to test $H_0$: $\lambda = \lambda_0$ vs. $H_1$: $\lambda \neq \lambda_0$.

c. Using your answer to b. write down the generalized likelihood ratio test statistic (LRT).

*Hint: Your answer should be a function of $n$, $\lambda_0$ and $\hat{\lambda}$.*

$$\begin{aligned}
\lambda_{lrt}(y_1 \ldots y_n) &= \frac{f(y_1 \ldots y_n|\lambda_0)}{\sup_{\lambda \in \mathbb{R}} f(y_1 \ldots y_n|\lambda)} \\
&= \frac{\lambda^{\sum_{i=1}^{n} y_i} e_0^{-n\lambda}}{\hat{\lambda}^{\sum_{i=1}^{n} y_i} e^{-n\hat{\lambda}}} \\
&= \frac{\lambda^{\sum_{i=1}^{n} y_i} e_0^{-n\lambda}}{\bar{y}^{\sum_{i=1}^{n} y_i} e^{-n\bar{y}}} \\
&= e^{-n(\lambda_0 - \bar{y})}\left(\frac{\lambda_0}{\bar{y}}\right)^{n\bar{y}}
\end{aligned}$$

d. Consider Bortkiewicz's Prussian Cavalry horse-kick fatality data:

| No. of fatalities | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| No. of years | 109 | 65 | 22 | 3 | 1 |

The original table from Bortkiewicz's 1898 book *Das Gesetz der kleinen Zahlen* is: `https://archive.org/stream/dasgesetzderklei00bortrich#page/n66/mode/1up`

Find the value of $\sum_{i=1}^{n} y_i$, and use your answer to b. to find the value of the MLE $\hat{\lambda}$.

*Hint: The table contains data on* $n = 200$ *observations,* $y_1, \ldots, y_{200}$*, of which* 109 *were* 0*; there were* 65 *which were* 1 *and so on.*

$$\sum_{i=1}^{200} y_i = 0 \cdot 109 + 65 + 2 \cdot 22 + 9 + 4 = 122$$

So the $\bar{y} = \frac{122}{200} \approx 0.6$

e. Use your answer to d. to find the LRT statistic for the hypothesis test with $\lambda_0 = 1$.

$$\begin{aligned}
\lambda_{lrt}(y_1 \ldots y_n) &= e^{-n(\lambda_0 - \bar{y})} \left(\frac{\lambda_0}{\bar{y}}\right)^{n\bar{y}} \\
&= e^{-200(1-\frac{122}{200})} \left(\frac{200}{122}\right)^{200\frac{122}{200}} \\
&= e^{-80} \left(\frac{200}{122}\right)^{122} \\
&= 2.064 \cdot 10^{-8}
\end{aligned}$$

f Report the approximate p-value.

*Hint: calculate* $-2\log$*(LRT), and compare to the appropriate* $\chi^2$ *distribution. See Lecture 10, slide 43. Recall that small values of the LRT correspond to evidence against* $H_0$*.*

## Problem 4

Suppose that we are planning an experiment to test hypotheses about the mean of a population that is known to be normal with standard deviation $\sigma = 4$. We wish to test the null hypothesis $H_0 : \mu = 0$ vs. the alternative $H_A : \mu > 0$. We intend to use a likelihood ratio test with significance level $\alpha = 0.05$.

a. For which values of $\bar{X}$, the sample mean, will we reject the null hypothesis. Express your answer as a function of sample size, $n$.

We are given that $\bar{X} \sim N(\mu, \sigma)$. Rewriting this to standard normal yields:

$$Y = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

Under the null hypothesis $\mu = 0$ so we relate $\alpha$ to $c$ by:

$$\begin{aligned}
\alpha &= P(\bar{X} > c | \mu = 0) \\
&= P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0\right) \\
&= P\left(Y > \frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0\right) \\
&= 1 - P\left(Y < \frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0\right) \\
1 - \alpha &= \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0\right) \\
0.95 &= \Phi\left(\sqrt{n}\frac{c - \mu}{\sigma} \middle| \mu = 0\right)
\end{aligned}$$

The CDF of a standard normal distribution equals 0.95 at the standard score of 1.644854 so we can determine $c$ to be:

$$\begin{aligned}
\sqrt{n}\frac{c}{4} &= 1.6644854 \\
c &= \frac{6.579416}{\sqrt{n}}
\end{aligned}$$

b. Suppose that we plan to obtain a sample of size 36. The researcher thinks that if the alternative is true then perhaps $\mu = 0.3$. Calculate the power of the test to reject the null hypothesis under this particular alternative hypothesis.

Equivalently to the previous problem we can start by writing:

$$\begin{aligned}
1 - \beta &= 1 - P\left(Y < \frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0.3\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma} \middle| \mu = 0.3\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}\left(\frac{6.579416}{\sqrt{n}} - \mu\right)}{\sigma} \middle| \mu = 0.3\right) \\
&= 1 - \Phi\left(\frac{6\left(\frac{6.579416}{6} - 0.3\right)}{4}\right) \\
&= 1 - \Phi(1.194854) \\
&= 0.116
\end{aligned}$$

c. Continuing from b., the scientist who is planning the experiment wishes to have power at least 90%. (Thus your calculation in b. shows that more than 36 observations are required.) Find approximately the smallest sample size at which this power can be achieved, against the specific alternative $\mu = 0.3$.

*Hint: repeat the calculation performed in* b. *at different sample sizes using trial and error: you may wish to use* `R` *or a spreadsheet to speed up this calculation.*

```
n = 1528, Power = 0.9000476355273681
```

```python
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
def problem4c():
    mu, sigma = 0.3, 4
    n = 36

    xi = lambda ni: sigma * 1.6499/np.sqrt(ni)
    power = lambda n_i: 1 - stats.norm.cdf(xi(n_i), loc=mu, scale=sigma/np.sqrt(n_i))

    while power(n) < 0.9:
        n += 1

    print(f"n = {n}, Power = {power(n)}")#, end='\r')
```

d. Suppose that the researcher obtains 200 samples, and observes $\bar{x} = 0.65$. Compute the p-value for this hypothesis test.

$$
\begin{aligned}
p_v &= P(\bar{X} > \bar{x} | H_0) \\
&= 1 - \Phi\left(\frac{\sqrt{200}(0.65 - 0)}{4}\right) \\
&= 0.0108
\end{aligned}
$$

## Problem 5

A researcher performs a sequence of independent experiments, up to and including the first 'success', after which the researcher stops. Each experiment has the same probability $p$ of success. Let $T$ be the number of experiments performed (including the first observed success). The researcher wishes to test the null hypothesis

a. $H_0$: $p = 0.25$,

b. against the alternative hypothesis

c. $H_1$: $p > 0.25$.

The researcher proposes to reject the null hypothesis if $T < 4$.

a. What is the significance level $\alpha$ of the test proposed by the researcher?

$$\begin{aligned}\alpha &= P(T < 4|H_0)\\ &= P(T \leq 3|H_0)\\ &= \sum_{t=1}^{3} p_0(1-p_0)^{t-1}\\ &= \left(\frac{1}{4} + \frac{3}{16} + \frac{9}{64}\right)\\ &= \frac{37}{64}\end{aligned}$$

b. What is the power of the researcher's test against the specific alternative hypothesis that $p = 0.5$.

$$\begin{aligned}1-\beta &= P(T < 4|H_1)\\ &= P(T \leq 3|H_1)\\ &= \sum_{t=1}^{3} p_1(1-p_1)^{t-1}\\ &= \sum_{t=1}^{3} 0.5^t\\ &= \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right)\\ &= \frac{7}{8}\end{aligned}$$

c. Re-express the researcher's rule for rejecting the null hypothesis in terms of the likelihood ratio:

$$\text{LRT} = p(t \mid p = 0.25)/p(t \mid p = 0.5).$$

Specifically, find the value $\ell$ such that the researcher will reject $H_0$ $p = 0.25$ in favor of $p = 0.5$ if LRT$< \ell$. (Note: There will be a range of values for $\ell$ that will give the same test.)

$$\begin{aligned}\Lambda(t) &= \frac{P(T = t|H_0}{P(T = t|H_1)}\\ &= \frac{p_0(1-p_0)^{t-1}}{p_1(1-p_1)^{t-1}}\\ &= \frac{1}{2}\left(\frac{3}{2}\right)^{t-1}\end{aligned}$$

*Hint: (For all parts) Geometric distribution!*

## Problem 6: Confidence Intervals

Consider a 95% confidence interval for the mean height $\mu$ in a population. Which of the following are true or false:

a. Before taking our sample, the probability of the resulting 95% confidence interval containing $\mu$ is 0.95.

   True.

b. If we take a sample and compute a 95% confidence interval for $\mu$ to be $[1.2, 3.7]$ then $P(\mu \in [1.2, 3.7]) = 0.95$.

   False. We can only think of $\mu$ as being random before taking a sample. Afterwards, it's just a number.

c. Before taking our sample, the center of a 95% confidence interval for the population mean is a random variable.

   True, before taking a sample the center of our interval is a random variable because it's based on the sampling a random variable.

d. 95% of individuals in the population have heights that lie in the 95% confidence interval for $\mu$.

   False. CI tells us the confidence that after drawing a random sample it's mean will lie in that range, not the other way around.

e. Over hypothetical replications out of one hundred 95% confidence intervals for $\mu$, on average 95 will contain $\mu$.

   True.

f. After obtaining our sample, the resulting confidence interval either does or does not contain $\mu$.

   True, we don't know if our sampled CI actually contains or doesn't contain the mean.

## Problem 7

Goldberger Qu. 11.6 (Assume that the observations are drawn from a Normal Distribution and see Lecture 10, slide 46.)