Sample Answer Lab10: Estimation and Confidence Intervals

Task 1: Study example 14-1 on pages 539-540 in BBR (0.5 points). Based on the observed spatial point pattern in Figure 1:

[a] Write into the figure by hand the observed number of points in each grid cell. If a point is right on the boundary, then assign it to just one grid cell. (0.1 points)

6	3	4	3
1	1	3	2
3	1	3	4
7	2	1	4

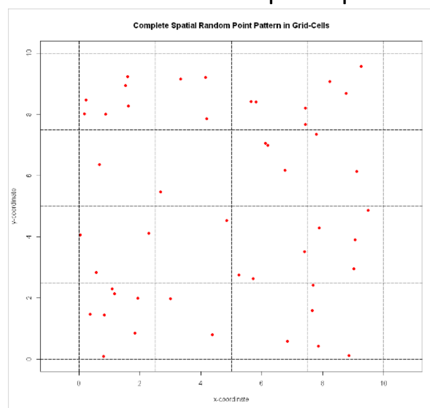


Figure 1: Random Point Pattern

[b] Manually estimate the ***mean*** number of points per grid cell (i.e., the average intensity). (0.2 points)

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{25} x_i}{n} \\ &= \frac{1*4+2*2+3*5+4*3+6*1+7*1}{16} \\ &= \frac{48}{16} = 3\end{aligned}$$

[c] Manually estimate the ***variance*** based on the observed number of the points per grid cell. (0.2 points)

$$\begin{aligned}S^2 &= \frac{\sum_{i=1}^{25}(x_i-\bar{X})^2}{n-1} \\ &= \frac{(1-3)^2*4+(2-3)^2*2+(3-3)^2*5+(4-3)^2*3+(6-3)^2*1+(7-3)^2*1}{16-1} \\ &= \frac{46}{15} = 3.0667\end{aligned}$$

Task 2: Confidence intervals (0.75 points)

[a] Show how a confidence interval around an unknown population expectation μ for an observed sample mean $\bar{x} = 100$ based on $n = 100$ sample observations from a normal distributed population with a standard deviation of $\sigma = 4$ at a given confidence level of 95% (or $\alpha = 0.05$) is calculate? Use the equation editor. (0.25 points)

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < z_{\alpha/2}$$
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$100 - 1.96 * \frac{4}{\sqrt{100}} < \mu < 100 + 1.96 * \frac{4}{\sqrt{100}}$$
$$99.216 < \mu < 100.784$$

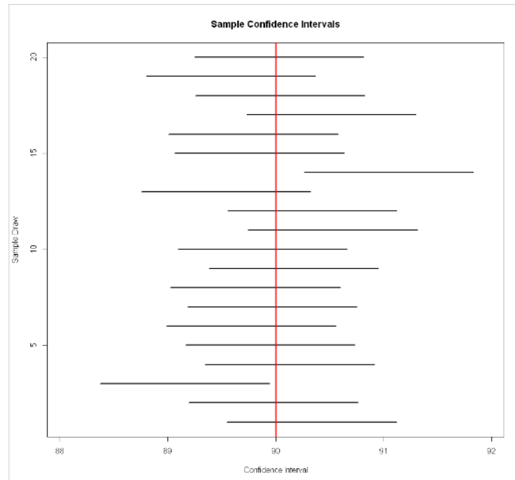


Figure 2: Confidence Intervals

[b] A set of 20 confidence intervals from 20 samples of size $n = 100$ with a population distribution of $X \sim N(\mu = 90, \sigma^2 = 16)$ has been calculated with an error probability of $\alpha = 0.1$. How many confidence intervals do you expect to ***not*** cover the population expectation $\mu = 90$ and ***explain*** why? (0.25 points)

Comment: A confidence interval, based on a sample from an unknown population, will cover with $1 - \alpha$ probability the true underlying population parameter. Therefore, for a 95% confidence interval we can expect that under repeated random sampling from an underlying population with an expectation of $\mu = 100$, in total 95% of the intervals will cover the underlying population expectation. In other words, for the given example with 20 samples, only 1 (5%) out of 20 sample confidence intervals will not cover the true population expectation.

[c] Does your sample from task 2 [a] support the assumption with 95 % probability that it was drawn from a population with $\sim N(\mu = 90, \sigma^2 = 16)$? (0.25 points)

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < z_{\alpha/2}$$
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$90 - 1.96 * \frac{4}{\sqrt{100}} < \mu < 90 + 1.96 * \frac{4}{\sqrt{100}}$$
$$89.216 < \mu < 90.784$$

Comment: No. The sample from task 2 [a] does not support the assumption with 95% probability that it was drawn from a population.

Task 3: A sample has been drawn from a normal distributed population. The estimated population mean is $\bar{x} = 8$ and the ***a priori known*** population variance is $\sigma^2 = 16$. (1 point)

Calculate the $1 - \alpha$ confidence intervals based on the parameters given in the table below. Insert the confidence intervals into the table below.

	$\alpha = 0.10$	$\alpha = 0.01$
$n = 16$	[6.35,9.65]	[5.42,10.58]
$n = 64$	[7.36,8.64]	[6.71,9.29]

Task 4: What trend with regards to the width of confidence intervals do you observe for decreasing error probabilities? (0.25 points)

Comment: With a decreasing error probability α, the confidence interval gets wider.

Task 5: What trend with regards to the width of confidence intervals do you observe for increasing sample sizes n? (0.25 points)

Comment: With an increasing sample size n, the confidence interval gets narrower because the standard error shrinks.

Task 6: You need to sample from a binomial distributed population to estimate the underlying population success probability π. You would like to be 95% confident that your estimation error is not larger than 0.1. How large does your sample need to be? Show your calculations typeset in a professional way. (0.25 points)

Comment: The population proportion is needed to evaluate the standard error in the equation for the sample size of the proportion estimator. Since the true population proportion is absolutely unknown prior to drawing a sample, a reasonable estimate needs to be used. In this case, it is advisable to adopt the **worst** case scenario for which the standard error is the largest. Therefore, we would assume $\pi = 0.5$.

$$E = 0.1$$

$$\alpha = 1 - 95\% = 5\%$$

$$z_{\alpha/2} = 1.96$$

When we use $p = 0.5$, we overestimate the sample size and therefore are on the safe side.

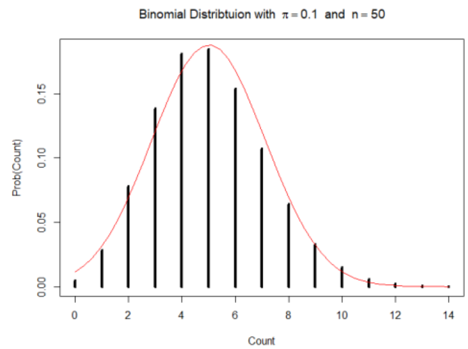
$$\begin{aligned} n &= \left[\frac{z_{\alpha/2} \cdot \sqrt{p(1-p)}}{E}\right]^2 \\ &= \left[\frac{1.96 \cdot \sqrt{(0.5)(0.5)}}{0.1}\right]^2 \\ &= 96.04 \\ &\approx 97 \end{aligned}$$

Task 7: The R script **`BinomToNorm.r`** plots the distribution of binomially distributed random counts X given the number of trials n and probability of success π. The density of normal distribution with an expectation $E[X] = n \cdot \pi$ and a variance $V[X] = \pi \cdot (1-\pi) \cdot n$ is superimposed onto the binomial distribution to visually compare whether it is approximately normal distributed. (0.6 points)

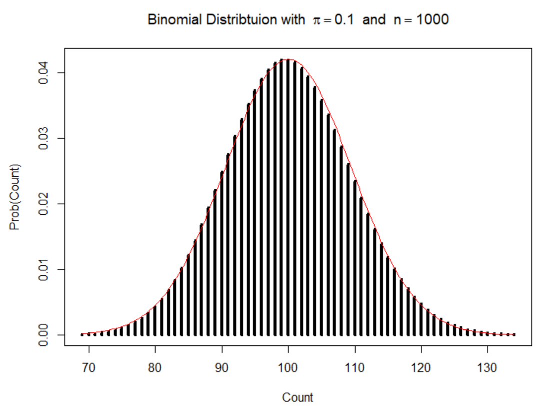
[a] Describe the shape of the binomial distribution for $n = 50$ (line 3) and $\pi = 0.1$ (line 4) and show the plot. Pay attention to the value range shown in the plot. (0.2 points)

```
n <- 50
pi <- 0.1
E <- pi * n
sdBinom <- sqrt(pi * (1 - pi) * n)
(skewBinom <- (1 - 2 * pi) / sdBinom)
[1] 0.3771236
```

Comment: The Shape of the binomial distribution for n = 50 and $\pi = 0.1$ is positive skewed.



[b] Increase the number of trials n to 1,000 (line 3) for an underlying probability of success $\pi = 0.1$ (line 4) and plot it. (0.2 points)



[c] ***In your judgment***, did binomial distribution sufficiently approach the normal distribution for the increased sample size? Perhaps argue in terms of the skewness (line 7) of both distributions. (0.2 points)

```
 > ( skewBinom <- (1-2*pi)/sdBinom )
[1] 0.0843274
```

Comment: As *n* increases to 1,000, the underlying probability distribution has sufficiently approached the normal distribution and the skewness of the distribution is close to zero. There is no substantial difference between the normal curve and the probabilities. The counts with a probability different from zero vary between 69 and 134 and are centered around the expected value of 900.

Task 8: Discuss the confidence intervals of the calculated regression parameter (0.4 points).

Use the model:

```
data(Cars93, package="MASS")
Cars93$MPG <- 0.7*Cars93$MPG.city+0.3*Cars93$MPG.highway
lm.cars <- lm(MPG~Fuel.tank.capacity+Weight+Passengers, data=Cars93)
summary(lm.cars)
```

[a] Evaluate the confidence intervals at a confidence level of 0.95 ($\alpha = 0.05$).

```
> cbind(coef=coef(lm.cars),confint(lm.cars, level=0.95))
```

	coef	2.5 %	97.5 %
(Intercept)	48.72965138	45.18445361	52.274849150
Fuel.tank.capacity	-0.49728784	-0.90492834	-0.089647346
Weight	-0.00552482	-0.00792306	-0.003126579
Passengers	0.18017363	-0.51185526	0.872202524

[b] Evaluate the confidence intervals at a confidence level of 0.99 ($\alpha = 0.01$).

```
> cbind(coef=coef(lm.cars),confint(lm.cars, level=0.99))
```

	coef	0.5 %	99.5 %
(Intercept)	48.72965138	44.033232421	53.426070336
Fuel.tank.capacity	-0.49728784	-1.037300183	0.042724501
Weight	-0.00552482	-0.008701834	-0.002347806
Passengers	0.18017363	-0.736575672	1.096922941

Why does the interpretation for the regression coefficient of the variable **`Fuel.tank.capacity`** change?

When $\alpha = 0.05$, the confidence interval for **`Fuel.tank.capacity`** variable is [-0.90492834, -0.0896473346], the estimated coefficient, -0.49728784 is within the interval. The true population slope differs with 95% probability from the non-relevant zero slope, therefore this variable is significant. However, when α decreases to 0.01, the upper bound of confidence interval increases to 0.042724501, now the interval covers zero, means that the variable may be not significant (coefficient is zero).